informatics

MDPI

Article

Multimodal Large Language Models vs. Human Authors: A Comparative Study of Chinese Fairy Tales for Young Children

Jing Du [1], Wenhao Liu [2], Dibin Zhou [2,3], Seongku Hong [1] and Fuchang Liu [2,*]

[1] Department of Media & Communication, Kangwon National University, Chuncheon 24341, Republic of Korea; dj@kangwon.ac.kr (J.D.); net9@kangwon.ac.kr (S.H.)
[2] School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China; whl819_819@163.com (W.L.); dibinz@zju.edu.cn (D.Z.)
[3] Center for Engineering and Scientific Computation, Zhejiang University, Hangzhou 310058, China
* Correspondence: liufc@hznu.edu.cn

Abstract

In the realm of children's education, multimodal large language models (MLLMs) are already being utilized to create educational materials for young learners. But how significant are the differences between image-based fairy tales generated by MLLMs and those crafted by human authors? This paper addresses this question through the design of multi-dimensional human evaluation and actual questionnaire surveys. Specifically, we conducted studies on evaluating MLLM-generated stories and distinguishing them from human-written stories involving 50 undergraduate students in education-related majors, 30 first-grade students, 81 second-grade students, and 103 parents. The findings reveal that most undergraduate students with an educational background, elementary school students, and parents perceive stories generated by MLLMs as being highly similar to those written by humans. Through the evaluation of primary school students and vocabulary analysis, it is further shown that, unlike human-authored stories, which tend to exceed the vocabulary level of young students, MLLM-generated stories are able to control vocabulary complexity and are also very interesting for young readers. Based on the results of the above experiments, we further discuss the following question: Can MLLMs assist or even replace humans in writing Chinese children's fairy tales based on pictures for young children? We approached this question from both a technical perspective and a user perspective.

Keywords: multimodal large language model; Chinese young children; fairy tales; story generation

check for updates

Academic Editor: Antony Bryant

Received: 18 October 2025
Revised: 24 November 2025
Accepted: 1 December 2025
Published: 9 December 2025

Citation: Du, J.; Liu, W.; Zhou, D.; Hong, S.; Liu, F. Multimodal Large Language Models vs. Human Authors: A Comparative Study of Chinese Fairy Tales for Young Children. *Informatics* **2025**, *12*, 139. https://doi.org/10.3390/informatics12040139

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The creation of original children's literature has always been a challenging aspect of Chinese children's literature. Children's fairy tales are universal, timeless, healing, and wonderfully magical, which is why they have endured throughout history [1]. The significance of a fairy tale lies in its ability to "shape" children into being virtuous, imaginative, and courageous [2]. Contemporary Chinese fairy tales have struggled to produce enduring works due to sociocultural shifts, limited creative infrastructure, and over-instrumentalization—often serving educational or commercial purposes rather than exploring universal themes [3]. Misalignment with authentic child perspectives and weak integration of indigenous cultural elements further limit their depth and resonance [4]. Research shows that children's fairy tales play a crucial role in various aspects of child

development, including stimulating imagination, aiding cognitive and language development, fostering emotional and moral growth, cultivating an interest in reading, enhancing concentration, conveying and promoting cultural traditions and values, improving logical thinking and analytical skills, and reducing stress [5]. Albert Einstein once said: "If you want your children to be intelligent, read them fairytales. If you want them to be more intelligent, read them more fairytales." [6] Can artificial intelligence (AI) be utilized to assist in the creation of children's literature? Recently, story generation has made significant progress with the help of large language models (LLMs), reaching a level very close to that of stories written by human authors in many respects. Artificial intelligence presents new possibilities to address these challenges. By analyzing global narrative structures and cultural motifs, AI can support plot innovation and cross-cultural creativity. It also enables age-appropriate language refinement and facilitates immersive storytelling through interactive technologies. Rather than replacing human creativity, AI serves as a tool to liberate authors from technical constraints, allowing them to focus on meaningful local themes—such as childhood experiences across rural and urban settings—and to craft stories that combine emotional depth with cultural authenticity.

Therefore, if large language models can assist human authors in generating high-quality children's fairy tales, they can significantly contribute to children's language education and aesthetic development. The use of LLMs such as GPT-4 [7] for creating various types of stories has achieved a level of quality very close to that of human authors [8]. FairyLandAI [9] creates personalized fairy tales for children, representing a pioneering step in leveraging LLMs for educational and cultural enrichment. Storypark [10] leverages LLMs to design an interactive storytelling system that offers children plot frameworks and interpretations of central themes throughout the storytelling process. This approach enhances children's expressive abilities, fosters creative thinking, and deepens their understanding of stories. However, most research and evaluation in this area have been conducted based on English stories [11]. As far as we know, there is relatively little research on LLMs for Chinese stories [12], especially Chinese children's fairy tale generation.

To the best of our knowledge, our study is the first to systematically explore this area. This paper makes two primary contributions. First, it offers a methodological advancement. Unlike prior studies that rely exclusively on benchmark metrics, we collected a dataset of human-authored stories from children's fairy tale websites for analytical purposes. We conducted comparisons between human-written stories and MLLM-generated stories not only in terms of linguistic attributes but also through vocabulary size analysis. This methodological approach more closely aligns with real-world parental usage scenarios and children's actual reading competencies. Second, our evaluation incorporates both student and parental acceptance. Our assessment framework includes evaluations from both student and parental perspectives, and through multi-dimensional analysis, it reveals parents' quality assessments, attitudinal orientations, and willingness to recommend AI-generated stories.

2. Literature Review

2.1. Evaluation on LLMs

LLMs are gaining increasing popularity in both academia and industry due to their unprecedented performance in various applications. As LLMs continue to play a vital role in research and everyday use, evaluating them becomes increasingly critical. This evaluation is essential not only at the task level but also at the society level to better understand their potential risks.

Evaluating LLMs is crucial for several reasons. First, it helps humans understand their strengths and weaknesses. Second, effective evaluation can guide the development

of human–machine collaboration. Third, extensive evaluation can enhance the safety and reliability of LLMs. Existing tasks for evaluating LLMs typically include the following: natural language processing, robustness, ethics, biases and trustworthiness, social sciences, natural sciences and engineering, medical applications, and agent applications (using LLMs as agents).

The initial objective behind the development of language models, particularly large language models, was to enhance performance on natural language processing tasks, including both understanding and generation. Natural language understanding encompasses a broad range of tasks aimed at achieving a deeper comprehension of the input sequence [13]. Reasoning tasks pose significant challenges for intelligent AI models [14], as they require not only comprehension of the provided information but also the ability to utilize reasoning and inference to deduce answers when explicit responses are absent. Natural language generation, on the other hand, evaluates the capabilities of LLMs in generating specific texts [15]. This includes tasks such as summarization, dialogue generation, machine translation, question answering, and other open-ended generation tasks [16]. Therefore, some LLM evaluation datasets are used to test and compare the performance of different language models on these aforementioned tasks, such as GLUE [17] and SuperGLUE [18]. However, most evaluation datasets are designed for the English language. For example, out of the 46 datasets mentioned in this survey [19], only 4 are for Chinese language evaluation.

Two common evaluation methods, automatic evaluation and human evaluation, are typically used to assess LLMs. Automatic evaluation includes several metrics [19]:

- Accuracy: Exact match, quasi-exact match, F1 score, ROUGE score.
- Calibrations: Expected calibration error, area under the curve.
- Fairness: Demographic parity difference, equalized odds difference.
- Robustness: Attack success rate, performance drop rate.

The increasingly advanced capabilities of LLMs have certainly surpassed standard evaluation metrics for general natural language tasks. Therefore, human evaluation becomes a natural choice in some non-standard cases where automatic evaluation is not suitable. For instance, in open-generation tasks where embedded similarity metrics (such as BERTScore [20]) are insufficient, human evaluation is more reliable [21].

Human evaluation includes several key aspects [19]:

- Accuracy: Assessing whether the information aligns with factual knowledge and avoids errors and inaccuracies.
- Relevance: Evaluating the appropriateness and significance of the generated content.
- Fluency: Examining the language model's ability to produce content that flows smoothly, maintaining a consistent tone and style.
- Transparency: Determining how well the model communicates its thought processes, enabling users to understand how and why certain responses are generated.
- Safety: Ensuring the language model avoids producing content that may be inappropriate, offensive, or harmful, thus protecting the well-being of users and preventing misinformation.
- Human alignment: Measuring the degree to which the language model's output aligns with human values, preferences, and expectations.

The expertise level of evaluators is a critical consideration, encompassing relevant domain knowledge, task familiarity, and methodological training. Ensuring that evaluators possess the necessary background knowledge to accurately comprehend and assess the domain-specific text generated by LLMs is essential. Therefore, the design of human evaluation is particularly important and should align with the evaluators' professional background knowledge, common sense, and capabilities. The above metrics primarily

focus on text generation tasks by large language models. For multimodal tasks involving image-based text generation, we adjusted and expanded these metrics. We retained Fluency and Relevance, expanded Accuracy into Coherence and Logic, and introduced Enjoyment as a new dimension specifically for children's fairy tales. We applied the five dimensions—Fluency, Relevance, Coherence, Logic, and Enjoyment—in the subsequent evaluation of both MLLM-generated and human-written stories.

2.2. Story Generation

Compared to more constrained text generation tasks like machine translation and summarization, which rely on existing content, story text generation is open-ended. It demands diversity and creativity while maintaining a continuous narrative. Automatic story generation is a classic task in the field of natural language processing (NLP) [22]. Since the advent of deep neural networks, especially with the development of larger parameter models and improved architectures, story generation has made significant progress. Unlike other natural language generation tasks such as machine translation, story generation tests the model's ability to apply the knowledge learned from training stories while maintaining coherence and consistency. Existing story generation models or systems have achieved good results in coherence and consistency, such as the autoregressive language models built on the Transformer architecture proposed by [23], including GPT [24], GPT-2 [25], BART [26], and GPT-3 [27]. Large language models, such as GPT-4, can produce stories that are significantly more coherent and fluent than those generated by state-of-the-art models from even a few years ago, such as GraphPlan [28] and Plan-and-Write [29]. However, most of the work on automatic story generation has focused on stories for an adult audience. Children's stories, despite their critical role in early literacy and future success [30], have not received as much attention.

Automatic story generation for children increases the potential for broader impact by allowing for the personalization of stories, making them more relevant to each individual child. For instance, tailoring stories to a child's specific interests could enhance their engagement with reading, thereby improving their literacy skills. Another application could be teaching preschoolers specific target words through customized stories. This study [31] examines the ability of several current LLMs to generate age-appropriate, simplified stories for children. They observe that despite their growing capabilities, modern LLMs are unable to generate children's stories with age-appropriate simplicity, particularly when compared to human-written counterparts. Another study [32] examines the trustworthiness of children's stories generated by large language models. The findings suggest that LLMs still struggle to produce high-quality children's literature and may even generate content that is inappropriate or toxic for children.

2.3. Parents' Acceptance of AI-Based Storytelling

Technology acceptance encompasses a user's willingness to adopt a system and considers both its practical and social acceptability [33]. A system deemed practically acceptable may still face social resistance. Parental involvement [34] is a crucial component of educational activities, including shared reading between parents and children. Shared reading offers a valuable opportunity for connection and learning, fostering skills that are essential for developing later independent reading abilities [35]. Previous research on the acceptance of AI-generated stories and parental behaviors remains limited. Relevant studies include Karl F. (2021) [36], which explores parental acceptance of a children's storytelling robot through the lens of the Uncanny Valley of AI theory, focusing on whether parents accept the robotic storytelling medium itself rather than the story content. Another study by Sun (2021) [37] examines the effectiveness of AI-based storytelling technologies in real-

world contexts, particularly regarding how well these technologies meet parents' needs and perceptions, with parents expected to play an active role in guiding and interacting with these AI systems. In contrast to previous studies, our research not only examines the language quality of AI-generated children's fairy tales but also investigates parents' likelihood of choosing these stories and identifies factors that may influence their willingness to choose them.

3. Research Method

We conducted three experiments to examine readability, logic, and entertainment value, as well as the perceived content quality, measured by the similarity of language quality, enjoyment, and plot between human-written stories and MLLM-generated stories, and the attitude towards the generated stories, measured by popularity, appeal, and assistance. Our approach not only complements and enhances existing evaluations of AI large models but also contributes to a systematic understanding of human subjective attitudes toward texts generated by large language models.

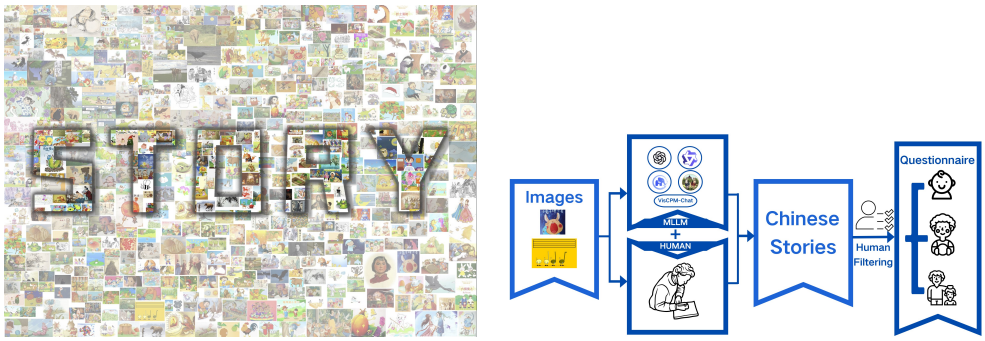
Based on the above focus, this paper presents the following hypothesis:

- H1: For evaluators with relevant majors, there is no significant difference between Chinese children's fairy tales written by humans and those generated by MLLMs for the same image prompt.
- H2: Elementary school students find stories generated by MLLMs to be as engaging as those written by humans.
- H3: Parents perceive stories generated by MLLMs to be acceptable and potentially usable for their children.

3.1. Study 1: Evaluating MLLM-Generated and Human-Written Stories

3.1.1. Participants and Materials for Study 1

As shown in Figure 1, we collected data and conducted our own surveys to perform significance testing and analyze the data to test our hypotheses. We collected 543 story images and their corresponding Chinese children's fairy tales from the internet (https://www.qigushi.com/ (accessed on 1 January 2025), https://www.gushi365.com/tonghuagushi/) (accessed on 1 January 2025). Subsequently, we employed five multimodal large language models (ChatGPT-4.0 [38], an optimized conversational system based on GPT-4, and four Chinese open-source MLLMs, including VisCPM [39], VCLA [40], VisualGLM [41], and QWen-VL [42]) to generate a unique Chinese fairy tale for each image. We first recruited three students from the fields of Chinese Language and Literature, Elementary Education, and Early Childhood Education to conduct a quality assessment of 543 groups of fairy tales. To ensure the suitability of the test materials, we adopted a two-stage filtering process to select the final set of stories. In the first stage, we performed screening based on image quality and relevance: images that were too small, of low resolution, or thematically inappropriate for children or fairy tale contexts were excluded. In the second stage, we evaluated the stories themselves, removing those that were either too short or excessively long, as well as those written in overly adult-oriented language. After this filtering process, we retained 50 stories and their corresponding images that met our criteria for content quality, appropriateness, and narrative balance. After the story selection process, we recruited a sample of 50 undergraduate students from the fields of Chinese Language and Literature, Elementary Education, and Early Childhood Education. Each participant was asked to rate one original story written by a human author, corresponding to a given image, and five additional stories generated by MLLMs based on the same image. We believe that undergraduate students, with their disciplinary background, are well-equipped to evaluate AI-generated stories from multiple dimensions.



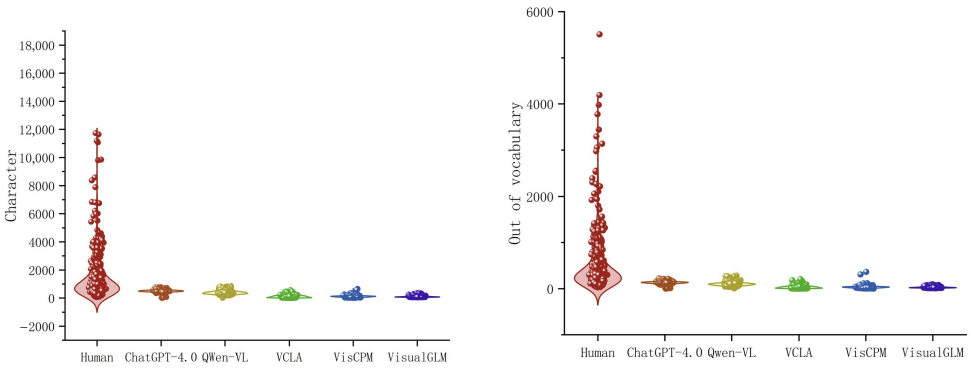
(**a**) Images collection of children's fairy tales. (**b**) Flowchart of our research.

Figure 1. Data preparation and research framework. (**a**) This is our story image dataset, consisting of 543 story pictures, which were scraped from Chinese fairy tale websites. (**b**) This is a pipeline of our research process, including story generation, human filtering, questionnaires, and statistical analysis.

3.1.2. Design and Procedure for Study 1

Analysis of generated story. Most studies testing large language models overlook the issue of vocabulary complexity in practical educational applications. An important task of the Chinese language curriculum is to cultivate students' interest in writing and encourage them to express their thoughts in written form. Although lower-grade primary school students have a certain level of character recognition and can carefully observe pictures, their overall language organization skills remain relatively weak. Their expressions may lack clarity and accuracy, and they may even struggle with logical coherence, resulting in incomplete content. Therefore, Chinese language teachers often use picture-based writing exercises to help students transition smoothly from oral to written expression [43]. To accurately assess whether large language models are suitable for young children, we consulted with primary school teachers and proposed using vocabulary levels from elementary textbooks as a reference for testing. We analyzed the differences between human-written stories and those generated by large models by examining the number of Chinese characters and whether the vocabulary exceeded the level appropriate for lower primary school students. It is worth noting that when these MLLMs generate children's fairy tales based on images, the stories are often quite short, typically less than ten sentences, and there is a noticeable difference compared to human-written stories. Even ChatGPT-4.0 often generates shorter stories in Chinese compared to generating English fairy tales. We speculate that this may be due to the fact that the scale of Chinese open-source data is smaller than that of mainstream English datasets, and the difficulty of Chinese word segmentation is higher than that of English. We believe this issue will soon be addressed and resolved through engineering efforts. We attempted to eliminate these differences by modifying the prompts. We use the prompt (translated into English) "Please tell a relatively long Chinese children's fairy tale based on this picture, which includes several paragraphs." to generate Chinese children's fairy tales. ChatGPT-4.0 and Qwen-VL-7B are able to generate relatively long Chinese children's stories. In contrast, VisCPM-10B, VCLA-7B, and VisualGLM-6B are mainly employed for question-answering (QA) tasks. Consequently, their abilities to create stories based on images are restricted, and thus the stories they generate are relatively short in length. However, human-written stories vary significantly in length, while stories generated by large language models tend to be more consistent in length, making them more suitable for younger children to read, as shown in Figure 2a. In terms of vocabulary, we referred to the vocabulary and character lists from the Chinese language textbooks for grades 1–2 published by the People's Education Press (Appendix A). Our findings reveal that human-authored fairy tales significantly exceed the vocabulary level of primary school students, whereas large language models can generate fairy tales suitable

for young children based on specific prompts, as shown in Figure 2b. In subsequent human evaluations, we aim to minimize the impact of these issues as much as possible. Therefore, in the subsequent study, we invited undergraduates majoring in early childhood education to assist us in selecting stories that have similar lengths and vocabulary to those generated by AI and human-written narratives for the questionnaire.



(**a**) (**b**)

Figure 2. Vocabulary analysis of human-written stories and stories generated by LLMs. (**a**) Comparison of the number of Chinese characters between human-written stories and stories generated by different LLMs. The character counts for the stories generated by ChatGPT-4.0, Qwen-VL, VCLA, VisCPM, and Visualglm, and those written by humans, are as follows: 511, 398, 126, 123, 124, and 1438. (**b**) Comparison of out-of-vocabulary Chinese word counts between human-written stories and stories generated by different LLMs. The word counts for the stories generated by ChatGPT-4.0, Qwen-VL, VCLA, VisCPM, and Visualglm, and those written by humans, are as follows: 135, 108, 36, 37, 33, and 492.

Readability evaluation. Each student was shown an image and then read six corresponding stories, one written by a human and five generated by different MLLMs. The students were then asked to answer five questions, rank the six stories based on their preference, and provide reasons for their top-two favorite stories. In the questionnaire, it was clearly indicated which story was written by a human and which ones were generated by MLLMs. The questionnaire was privately sent to each respondent via WeChat APP. This experiment was conducted in Chinese. Our human evaluation assesses MLLM-generated children's fairy tales from five dimensions: Fluency (how many grammatical errors are in the story), Coherence (whether there are inconsistencies or irrelevant details in the story), Relevance (how well the story description relates to the picture), Logic (how well the story description adheres to common sense), and Enjoyment (how much you liked the story after reading it). All scale evaluations were conducted using a 5-point Likert scale.

Question 1 (Fluency): One point represents a story with too many grammatical errors and is completely unreadable, while five points represent a story with no grammatical errors and clear, coherent text.

Question 2 (Coherence): One point represents a story with contradictory and unrelated events, while five points represent a story with no contradictions and fully consistent content.

Question 3 (Relevance): One point represents a story completely unrelated to the picture, while five points represent a story fully aligned with the picture content.

Question 4 (Logic): One point represents a story with no logical sense, while five points represent a story that fully adheres to common sense.

Question 5 (Enjoyment): One point represents a story that is uninteresting and makes the reader want to stop reading as soon as possible, while five points represent a story that is very interesting and makes the reader want to continue reading.

Question 6 (Rank): Rank the human-written story and the stories generated by the five models based on preference, with 1 point being the most liked and 5 point being the least liked.

3.2. Study 2: Distinguishing MLLM-Generated from Human-Written Stories

3.2.1. Participants and Materials for Study 2

We first surveyed 30 first-grade students and 81 second-grade students in an elementary school. The experiment was conducted during the long break time of the primary school, with homeroom teachers distributing the questionnaires. Each elementary school student read and evaluated the three stories based on the story images provided in the questionnaire. The elementary school students answered the questions independently, and the completed questionnaires were collected by the teachers. Prior to conducting experiments with elementary school students, we invited a Chinese language teacher from a primary school in Shaoxing (a city in Zhejiang province, China) to assist us in selecting stories. Considering the characteristics of lower-grade primary school students [44]: The lower grades of primary school represent a critical transition phase during which children shift from "acquiring oral language" to "learning written-form reading". Chinese characters are abundant in number and lack clear and systematic shape–sound conversion rules, rendering them strikingly different from alphabetic scripts. Children with subpar early reading skills are likely to derive less enjoyment from reading and show less inclination to participate in reading activities. Inadequate reading practice subsequently causes a continuous deterioration in their reading capabilities, thereby forming a vicious cycle. The basis for selecting the stories was the vocabulary list from the Chinese language textbooks for grades 1 to 2, as outlined in the People's Education Press curriculum. This experiment was conducted in Chinese. Then, we recruited 103 participants to answer a questionnaire hosted on the WJX.CN platform, all of whom were parents of students from the same school.

3.2.2. Design and Procedure for Study 2

Since the previous study informed participants whether a story was human-written to facilitate their comparison of human- and MLLM-generated stories, this study eliminates potential psychological bias by not disclosing the origin of the stories to the participants and asks them to identify which story is human-written. The effect of prompting on story generation by MLLMs is not the focus of this study. We therefore adopt a simple strategy by testing three prompts—one in English and two in Chinese. Examples of the stories generated by the five multimodal large language models, together with their corresponding prompts, are provided in Appendix B.

Evaluation by elementary school students. This experiment aims to replicate the results of Experiment 1, which targeted elementary school students. Through the analysis of Experiment 1, we found that the stories generated by ChatGPT-4.0 and QWen-VL-7B were significantly better in terms of text length and the five indicators (Question 1∼Question 5) compared to other models. Considering the reading speed of elementary school students, we simplified the 50 questionnaires in study 1, and only retained three stories for each questionnaire. These three stories include the story written by a human, as well as the stories generated by ChatGPT-4.0 and QWen-VL-7B. Considering the limited logical thinking ability of lower-grade elementary school students, we only used the Enjoyment evaluation metric and simplified it to a 3-point Likert scale, where 1 represents boring, 2 represents so-so, and 3 represents interesting. After reading, the students were asked to rank the three stories based on their preference, from most to least liked, and to write down the reasons for their top choice. Each student completed only one questionnaire.

Evaluation by parents. We conducted this experiment to examine readability, logic, and entertainment value, as well as the perceived content quality, measured by the similarity of language quality, enjoyment, and plot between human-written stories and MLLM-generated stories, and the attitude towards the generated stories, measured by popularity, appeal, and assistance. Favorite and Willingness reflect the degree of parents' preference for the generated stories and their willingness to recommend them. Each parent was required to complete only one questionnaire.

In order to make the stories generated by the LLM have the most similar appearance to human-written stories, we selected three groups of stories from the fifty groups of stories in study 1, each of which only contains two stories (the stories written by humans and the stories generated by ChatGPT-4.0 with the closest lengths), and randomly sent them to the parents. Each participant was required to answer eight questions, which are detailed as follows (all scale evaluations were conducted using a 5-point Likert scale):

Question 1 (Quality): Choose the similarity of the language quality of the two stories. One point represents very dissimilar, while five points represent very similar.

Question 2 (Enjoyment): Choose the similarity of the interestingness of the two stories. One point represents very dissimilar, while five points represent very similar.

Question 3 (Plot): Choose the similarity of the plot of the two stories. One point represents very dissimilar, while five points represent very similar.

Question 4 (Popularity): Do you think the stories generated by AI will be popular among child readers? One point represents very unpopular, while five points represent very popular.

Question 5 (Appeal): Do you think the stories generated by AI can attract children to read? One point represents very unattractive, while five points represent very attractive.

Question 6 (Assistance): Do you think the stories generated by AI are helpful for children's language learning? One point represents very unhelpful, while five points represent very helpful.

Question 7 (Favorite): Which story do you prefer? One point represents human-written stories, while two points represent AI-generated stories.

Question 8 (Willingness): Do you recommend that children read the stories generated by AI? One point represents highly not recommended, while five points represent highly recommended.

4. Results

4.1. Human Evaluation of Human-Written and MLLM-Generated Stories

Table 1 gives quantitative results of scores by undergraduate students. We used the ranks from Table 1 as scores to conduct independent samples *t*-tests to examine the significance of differences between human-written stories and those generated by ChatGPT 4.0 [$M_{\text{human}} = 1.86$, $SD_{\text{human}} = 1.34$, $M_{\text{ChatGPT}} = 2.48$, $SD_{\text{ChatGPT}} = 1.27$, $t(98) = 0.13$, $p = 0.72$, 95% CI $= [-1.14, -0.10]$, Cohen's d$= 1.30$], as well as between human-written stories and those generated by QWen-VL-7B [$M_{\text{human}} = 1.86$, $SD_{\text{human}} = 1.34$, $M_{\text{QWen}} = 2.88$, $SD_{\text{QWen}} = 1.33$, $t(98) = 0.20$, $p = 0.66$, 95% CI $= [-1.55, -0.49]$, Cohen's d$= 1.34$]. The results of ChatGPT 4.0 and QWen-VL-7B indicate that H1 is accepted. Similarly, we give the t-tests results of VisCPM-10B [$M_{\text{human}} = 1.86$, $SD_{\text{human}} = 1.34$, $M_{\text{VisCPM}} = 4.58$, $SD_{\text{VisCPM}} = 1.37$, $t(98) = -10.03$, $p < 0.01$, 95% CI $= [-3.26, -2.18]$, Cohen's d$= 1.36$] and VCLA-7B [$M_{\text{human}} = 1.86$, $SD_{\text{human}} = 1.34$, $M_{\text{VCLA}} = 4.76$, $SD_{\text{VCLA}} = 1.00$, $t(98) = -12.26$, $p < 0.01$, 95% CI $= [-3.37, -2.43]$, Cohen's d$= 1.18$] and VisualGLM-6B [$M_{\text{human}} = 1.86$, $SD_{\text{human}} = 1.34$, $M_{\text{VisualGLM}} = 4.44$, $SD_{\text{VisualGLM}} = 1.37$, $t(98) = -9.51$, $p < 0.01$, 95% CI $= [-3.12, -2.04]$, Cohen's d$= 1.36$]. The experiment indicates that most undergraduate students find the stories generated by ChatGPT-4.0 and QWen-VL-7B to

be very similar to those written by humans. The other three models have significantly different scores compared to human-written stories due to the generated stories being too short. The reason is that these models primarily focus on simple visual question-answering tasks. We also give the Scheffé Test results of story rankings by undergraduate students in Table 2, which support the preceding conclusions.

Most people selected the human-written story as their top choice and ChatGPT-4.0 as their second choice. The primary reason for selecting the human-written story as the top choice was that the language was elegant, the story details were specific, and it was realistic and aligned with common sense. The main reason for selecting the second choice was that the story was considered more creative and interesting. The users' reasons align with the data in Table 1, which also shows that human-written stories are more correlated with the images and adhere to common sense, whereas ChatGPT-generated stories are more interesting and coherent. Table 3 gives ANOVA results for students from three majors on their top-two favorite stories. First, we encoded the methods of story generation. The stories generated by a human, ChatGPT-4.0, VisCPM, VCLA, VisualGLM, and Qwen-VL were assigned codes from 1 to 6, respectively. As shown in Table 3, there are significant between-group differences in the selection of the top-one story ($p = 0.046$), while for the top-two story, the differences are not significant. That is, students from different majors basically converge in choosing LLMs for the top-two story, while for the top-one story, some tend to choose human-written stories and others tend to choose LLM-generated stories. Through the analysis of user reasons, it can be seen that students majoring in Chinese Language and Literature attach great importance to Fluency. Students majoring in Elementary Education focus more on the Logic of the story. Students majoring in Early Childhood Education attach importance to whether the story is suitable for children to read, thus considering various dimensions, while the top-two story mainly comes from ChatGPT-4.0 and Qwen-VL-7B, and the reasons for selection are mostly related to the aspect of enjoyment. These results are consistent with the Rank index in Table 1. Of course, this assessment was made under the condition that the survey participants were informed which stories were AI-generated, which may have introduced some bias. Participants might have been psychologically influenced, leading them to perceive the human-written story as better. Tables 4–8 present a detailed frequency list of the reasons provided by the undergraduate students.

Table 1. Quantitative results of scores by undergraduate students.

Method	Fluency ↑	Coherence ↑	Relevance ↑	Logic ↑	Enjoyment ↑	Rank ↓
Human	3.94	4.20	**3.94**	**4.10**	3.38	**1.86**
ChatGPT-4.0	**4.38**	**4.30**	3.56	4.02	**3.46**	2.48
VisCPM-10B	4.30	4.14	3.02	3.40	1.94	4.58
VCLA-7B	4.28	3.68	2.64	3.06	2.02	4.76
VisualGLM-6B	4.28	4.04	2.96	2.92	1.90	4.44
QWen-VL-7B	4.36	4.24	3.34	3.60	3.28	2.88

The bold indicates the best scores. ↑ indicates higher values are better.

Table 2. Scheffé Test results of story rankings by undergraduate students.

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% CI Lower Bound	95% CI Upper Bound
Human	ChatGPT-4.0	−0.62	0.26	0.329	−1.48	0.24
	VisCPM-10B	−2.72	0.26	<0.001	−3.58	−1.86
	VCLA-7B	−2.90	0.26	<0.001	−3.76	−2.04
	VisualGLM-6B	−2.58	0.26	<0.001	−3.44	−1.72
	QWen-VL-7B	−1.02	0.26	0.009	−1.88	−0.16

Table 3. ANOVA results for the top-two favorite stories among three majors. Group 0: Chinese Language and Literature, Group 1: Elementary Education, Group 2: Early Childhood Education.

	Group	Num	Mean	SD	95% CI	Sig.
top-one	0	6	3.33	2.58	[0.62, 6.04]	
	1	20	2.35	2.18	[1.33, 3.37]	
	2	24	1.46	0.93	[1.07, 1.85]	
	Total	50	2.04	1.83	[1.52, 2.56]	
						0.046 (Between Groups)
top-two	0	6	2.00	0.63	[1.34, 2.66]	
	1	20	2.80	1.74	[1.99, 3.61]	
	2	24	3.17	1.99	[2.33, 4.01]	
	Total	50	2.88	1.79	[2.37, 3.39]	
						0.357 (Between Groups)

Table 4. The frequency list of reasons given by the undergraduate students for choosing human stories.

Reason	Frequency
"Rich in content."	1
"Very suitable for children to read before going to bed. Mentioning sleep many times implies that children should go to sleep."	1
"Rich in language and more suitable for children to read."	1
"Familiar plot and interesting."	1
"Described in detail."	1
"Vivid and clearer description."	1
"Detailed plot."	1
"Her description is very suitable for children's reading."	1
"Suitable for children's reading."	1
"There is a plot and a revelation."	1
"Human thought and creativity are above artificial intelligence."	1
"Suitable for children's cognitive levels."	1
"Ups and downs in the plot."	1
"Wisdom with underlying philosophical insights."	1
"Quite reasonable, like the stories we often heard in childhood."	1
"Cows like to eat various fruits and can produce milk with various fruit flavors. It's interesting."	1
"More humane language."	1
"Interesting."	2
"Interesting. Conform to logic."	1
"Rich in plot."	1
"Realistic."	1
"Good imagination."	1
"Beautiful language and strong logical coherence."	1
"Somewhat absurd yet interesting.'	1
"The language is suitable for children's reading."	1
"The human stories have good logical coherence, use relatively simple language, and are suitable for young children to read."	1
"No reason."	6

Table 5. The frequency list of reasons given by the undergraduate students for choosing ChatGPT-4.0's stories.

Reason	Frequency
"Good imagination."	1
"The story is coherent, with rhetorical techniques and vivid language description. There is a sense of picture in detailed description."	1
"Logical, simple and clear language with profound value."	1
"The story is relatively smooth."	1
"More lively and vivid."	1
"The story plot is relatively reasonable, the language description is vivid and uses personification rhetorical techniques. The detailed description has a sense of picture."	1
"Innovative."	1

Table 6. The frequency list of reasons given by the undergraduate students for choosing VisCPM's stories.

Reason	Frequency
"Read smoothly and is also relatively interesting."	1

Table 7. The frequency list of reasons given by the undergraduate students for choosing VisualGLM's stories.

Reason	Frequency
"Cute."	1
"More story plots."	1

Table 8. The frequency list of reasons given by the undergraduate students for choosing QWen-VL's stories.

Reason	Frequency
"A snowman can build a snowman. That's so fun."	1
"Innovative"	1
"I like."	1
"Vivid"	1
"Full of imagination and not a dull statement of pictures."	1
"Full plot."	1
"No reason."	1

4.2. Response of Elementary School Students

We made sure that there were no differences between cases of first-grade students in which the stories were written by a human or ChatGPT 4.0 [$M_{\text{human}} = 1.87$, $SD_{\text{human}} = 0.82$, $M_{\text{ChatGPT}} = 2.10$, $SD_{\text{ChatGPT}} = 0.76$, $t(58) = 0.69$, $p = 0.41$] and by human or QWen-VL-7B [$M_{\text{human}} = 1.87$, $SD_{\text{human}} = 0.82$, $M_{\text{QWen}} = 2.00$, $SD_{\text{QWen}} = 0.87$, $t(58) = 0.13$, $p = 0.72$]. Similarly, we made sure that there were no differences between cases of second-grade students in which the stories were written by a human or ChatGPT 4.0 [$M_{\text{human}} = 1.84$, $SD_{\text{human}} = 0.81$, $M_{\text{ChatGPT}} = 2.00$, $SD_{\text{ChatGPT}} = 0.82$, $t(160) = 0.31$, $p = 0.58$] and by a human or QWen-VL-7B [$M_{\text{human}} = 1.84$, $SD_{\text{human}} = 0.81$, $M_{\text{QWen}} = 2.15$, $SD_{\text{QWen}} = 0.80$, $t(160) = 0.25$, $p = 0.62$]. These results indicate that, for elementary school students,

there is no significant difference between human-written stories and AI-generated stories. Therefore, the H2 hypothesis is accepted.

From Tables 9 and 10, it can be seen that the scores and ranks given by elementary school students for the human-written stories and the two MLLM-generated stories are very close. This indicates that elementary school students indeed find it difficult to distinguish between human-written stories and MLLM-generated stories. From the word cloud analysis in Figure 3, 30 first-grade students cited educational value as the primary reason for liking their favorite stories, with 12 students mentioning friendship and 12 students highlighting funny as key factors (Figure 3a). A total of 81 second-grade students provided different reasons, with 50 students stating that they liked the story because it was interesting, 32 students emphasizing its imaginative aspects, and 41 students appreciating its moral lesson (Figure 3b). This indicates that MLLMs can also generate educational stories that teach children some truths about the world.



(**a**) (**b**)

Figure 3. Word cloud analysis of reasons for the favorite story given by elementary students. (**a**) Reasons for first-grade students' favorite story. (**b**) Reasons for second-grade students' favorite story.

Table 9. Quantitative results of scores on the first-grade elementary school students.

Method	Enjoyment ↑	Rank ↓
Human	2.47	**1.87**
ChatGPT-4.0	2.47	2.10
QWen-VL	2.47	2.00

The bold indicates the best scores. ↑ indicates higher values are better.

Table 10. Quantitative results of scores on the second-grade elementary school students.

Method	Enjoyment ↑	Rank ↓
Human	**2.53**	**1.84**
ChatGPT-4.0	2.40	2.00
QWen-VL-7B	2.24	2.15

The bold indicates the best scores. ↑ indicates higher values are better.

4.3. Parents' Distinguishing Ability and Evaluation of MLLM-Generated Stories

Table 11 shows that most parent ratings are close to 4 (the maximum score is 5), indicating that they find the stories generated by large language models beneficial for children and are willing to recommend them for reading. In fact, using large language models to generate children's stories is an inevitable and unavoidable application, with many parents already adopting this approach to engage their children. Therefore, parents are generally receptive to artificial intelligence and its related products. Many young

parents in particular have already started using various AI-powered educational products to accompany their children in their leisure time. Another interesting finding is as follows: for Question 7, the favorite score is 1.62, indicating that out of 103 parents, 64 (or 62.14%) preferred the story generated by the large model. This indicates that most parents, without being informed which story was AI-generated, made their judgments about whether a story was AI-generated in a manner close to blind guessing. These findings indicate that the H3 hypothesis is supported. We examined the internal consistency reliability of each dimension of the scale. The story similarity includes similarity of language quality, similarity of enjoyment, and similarity of plot. The consistency of three similarity properties is good (Cronbach's $\alpha = 0.71$). The perceived attitude includes language learning assistance, popularity, and appeal. The consistency of three attitude properties is good (Cronbach's $\alpha = 0.76$).

Table 11. Quantitative results of scores on the parents of elementary school students.

Quality ↑	Enjoyment ↑	Plot ↑	Popularity ↑	Appeal ↑	Assistance ↑	Favorite ↓	Willingness ↑
3.93	3.83	4.01	3.93	3.89	4.00	1.62	3.78

5. Discussion

5.1. The Impacts of AI Tools Represented by LLMs on Education

As humanity has always done, we create and utilize machines to transform our lives and deepen our understanding of life itself [45]. From Jaquet-Droz's 18th-century writing automaton to modern large language models, technological advancements continue to shape our perceptions. Thinkers such as Descartes, Lamarck, and Schrödinger have contributed to the ever-evolving exploration of the relationship between humans and machines, as well as between matter and consciousness. AI has significantly transformed cultural creation, enabling the generation of paintings [46], music [47], and poetry [48], as well as the development of screenplays and theatrical scripts [49,50]. FairyLandAI [9] marks a significant advancement in AI-driven personalized storytelling, highlighting its potential to enhance children's educational and moral development through customized narrative experiences. Recently, numerous AI-powered educational products have emerged in the market. Some of these tools are mature commercial products (e.g., Alpha Egg [51], Luka [52], Codi [53]), while others are still in the research phase (e.g., StoryCoder [54] and StoryBuddy [55]). Although these tools fundamentally rely on existing AI models for story generation, the use of AI-based tools to support interactive storytelling between parents and children has become a contemporary trend.

However, from the perspective of technology users, effectively utilizing this technology is a complex task for educators and parents. For instance, can the generated content adapt to a child's growing vocabulary, and does it incorporate sufficient grammar elements, cultural, historical, and moral content? How parents and children engage with LLM-generated stories, as well as the dynamics between parents, children, and LLMs, warrants further investigation. While children's primary interaction partners have traditionally been parents and teachers, recent advancements in AI have led to the emergence of numerous AI-driven storytelling and reading technologies. As these technologies become increasingly integrated into the lives of preschool children, critical questions arise regarding their role in real-world storytelling and reading contexts, as well as how parents, as key stakeholders, perceive and experience these innovations. The authors of [37] present the use of AI based on two design principles: "Taking Parents as Beneficiaries: Support, not Substitute" and "Enabling Parents' Personal Experience Through Co-Creation".

From the perspective of primary school teachers, picture-based writing (tupian xiehua) is widely recognized as a foundational component of early literacy education—it cultivates students' observational skills, logical thinking, and basic expression abilities, serving as a critical bridge between visual perception and written communication. Teachers often emphasize that traditional picture-based writing instruction faces the challenge of overly uniform model essays, which may constrain students' creative thinking by encouraging mechanical imitation rather than independent expression. In this context, the stories created by MLLMs offer a distinct advantage: by generating diverse, contextually rich narrative frameworks tailored to different picture prompts, they can break the rigidity of standardized model essays and expose students to varied storytelling structures, thematic angles, and linguistic expressions. This diversity not only enriches students' learning resources but also inspires their creative potential, guiding them to explore unique ways of interpreting pictures and constructing narratives. However, primary school teachers also express legitimate concerns regarding the practical application of this approach. A key worry is that over-reliance on AI-created stories may lead students to passively accept ready-made content without engaging in deep thinking—such as analyzing picture details, organizing logical sequences, or refining language independently. This could undermine the core educational goal of picture-based writing, which prioritizes the development of students' own cognitive and expressive capacities. Consequently, teachers emphasize that parental supervision and guidance are indispensable in the implementation process: parents should encourage students to first independently analyze pictures and draft their own ideas before referencing AI-created stories, using the latter as a source of inspiration rather than a direct template. Additionally, teachers suggest that clear guidelines should be established to help students distinguish between "learning from AI" and "copying AI," ensuring that the technology serves as a supplementary tool to enhance learning outcomes rather than a replacement for active intellectual engagement.

5.2. The Influence of AIGC on Cultural Innovation and Inheritance

From the perspective of cultural innovation, the significance of LLMs in story generation currently lies in providing a structured approach to creation, enhancing content generation efficiency while ensuring personalization and customization [56]. These models can effectively mimic specific styles, assist in education, and improve learning efficiency. The practice of assigning LLMs distinct personalized roles and contextual attributes for content generation has become increasingly prevalent, greatly enhancing the customization and adaptability of generated content [57]. Poetry was once one of the few textual domains that generative AI language models struggled to replicate convincingly. However, recent research suggests that AI's capabilities have surpassed prior expectations [58]. AI-generated poetry tends to be more accessible, making it easier for readers to grasp its imagery, themes, and emotions, whereas human poetry is often more complex. As a result, human readers surprisingly show a preference for AI-generated poetry. However, LLM-generated content exhibits formulaic characteristics—relying on statistical patterns, pattern matching, and the reassembly of existing texts—so it fundamentally differs from human creativity, lacking genuine subjective intent and true innovation. Consequently, AI still falls short of fully replacing human artistic expression. However, using AI to assist in the generation of children's stories appears to be a viable approach, as AI-generated narratives can be more easily customized by teachers and parents to suit individual needs.

With the advent of recent artificial intelligence-generated content (AIGC) technology, its integration into education has enabled more personalized and adaptive teaching and learning experiences. AIGC also plays a dual role in cultural heritage and development, which facilitates cultural innovation and dissemination by generating new cultural works.

However, its application also raises concerns about cultural sustainability. Fairy tales generated by LLMs may lack deep cultural connotations, potentially affecting cultural continuity. LLM-generated stories present both challenges and opportunities for cultural transmission, so the key lies in designing and training AI to not only create efficiently but also respect and preserve cultural essence. Through human–AI collaboration, expert guidance, and high-quality data training, AIGC can promote cultural dissemination while avoiding superficiality, ultimately achieving genuine cultural sustainability [59].

5.3. Technological Breakthroughs and Ethical Considerations

Improvements have been achieved by models such as Gemini, ChatGPT, DeepSeek-V3, and Claude 3.5 through means [60] including multimodal expansion, extension of the context window, optimization of the attention mechanism, utilization of the mixture of experts, and reinforcement learning from human feedback. Nevertheless, the architecture of LLMs remains the Transformer architecture, with self-attention mechanism [23] at its core. The training of LLMs is still restricted by training data and computational power, and their capabilities still remain a significant distance from attaining Artificial General Intelligence (AGI).

Moreover, large language models are a valuable asset for all of humanity and should be applied across various domains of society. Therefore, experts from different fields must rigorously test and evaluate these models, as broader adoption across industries is crucial for identifying potential issues. In the educational domain, more research is needed to assess the performance of large models in different learning contexts. Many researchers [61–64] recognize the vast potential of AIGC in educational applications. However, its adoption faces challenges, including subjective biases, technical reliability and security issues, the adaptability of educational content, and cultural and ethical concerns, which hinder its broader implementation in the education sector [61]. Therefore, the key challenge lies in how to leverage AIGC technology to provide more precise and effective guiding recommendations for its educational applications. These challenges present promising directions for technological development. Notably, our framework creates stories through direct inference using open-source pre-trained models (e.g., Qwen series) and ChatGPT, which may present potential biases in the content due to implicit social stereotypes embedded in the models' training data—such as rigid associations between gender and occupations, cultural perspective monoculture, or one-sided depictions of minority groups. Such biases could exert additional impacts in educational settings: when used as teaching aids, extracurricular reading materials, or creative writing examples, they may reinforce students' inherent perceptions and undermine the inclusive educational value of the content. Fortunately, these issues are addressable through technical means. Many MLLMs already incorporate built-in restricted vocabulary mechanisms, making such biases surmountable with targeted adjustments.

5.4. Computational Cost Analysis of MLLM Story Generation

While the experimental results demonstrate the effectiveness of leveraging open-source MLLMs (e.g., Qwen series) and ChatGPT for story generation via direct inference calls, it is noteworthy that computational cost analysis of the proposed framework exhibits distinct characteristics due to the absence of model fine-tuning. Unlike fine-tuning paradigms that incur substantial computational overhead from parameter updates, gradient calculations, and multi-epoch training (often requiring high-end GPUs or multi-card configurations), our approach relies solely on off-the-shelf inference capabilities of pre-trained models. For open-source MLLMs deployed locally, the core computational cost during inference is dominated by inference latency and hardware resource consumption

(e.g., GPU memory usage, CPU utilization), which are primarily determined by model scale (e.g., 7B, 13B, 72B parameters), input context length, and generation token count. Notably, the "zero fine-tuning" design of our framework significantly reduces computational barriers for story generation tasks.

5.5. Limitations and Future Research

Due to resource limitations for conducting the questionnaire, we selected only a representative set of human-written fairy tales and AI-generated fairy tales for the study. Although we recruited 103 parents, potential biases remain, such as the parents' marital status, income levels, and degree of involvement in child-rearing. The tests with undergraduate and elementary school students had a relatively modest sample size, suggesting room for expansion in future studies. Furthermore, our study focused exclusively on fairy tales for young children and did not extend to evaluating stories for older age groups. The story types were also limited, primarily featuring animal-based tales, with fewer examples from genres like adventure, fantasy, and humor. Our work primarily focuses on fairy tales for younger children, particularly those below the second grade. We have not yet conducted in-depth research on fairy tales suitable for older children, as these require more advanced language and content. In future work, we will explore generating other types of children's stories, such as children's science fiction.

6. Conclusions

In this paper, we construct a novel benchmark dataset comprising 543 carefully selected image–story pairs tailored to Chinese children's fairy tales. This dataset addresses the notable lack of multimodal resources for Chinese story generation and serves as a valuable foundation for evaluating the storytelling capabilities of MLLMs in child-oriented contexts. Leveraging this dataset, we systematically assessed five open-source MLLMs using multi-dimensional human evaluation criteria. We further conducted user studies involving first- and second-grade primary school students, comparing human-authored and AI-generated stories in terms of comprehensibility and engagement. In parallel, we surveyed parents to understand their acceptance and perceived usefulness of AI-generated stories in educational scenarios. Our findings not only demonstrate the feasibility of using MLLMs for Chinese children's story generation but also provide practical insights into their future integration into early childhood language education.

Author Contributions: Conceptualization, F.L. and J.D.; methodology, J.D. and D.Z.; software, J.D.; validation, F.L.; formal analysis, W.L.; investigation, J.D.; resources, S.H.; data curation, J.D. and F.L.; writing—F.L. and J.D.; writing—review and editing, F.L.; visualization, J.D.; supervision, F.L.; project administration, F.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The collected data are publicly available at the following link: https://github.com/liufc2020/ChineseFairyStory4Children (accessed on 18 October 2025).

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

The link to the vocabulary and character lists from the Chinese language textbooks for grades 1–2, published by the People's Education Press: https://zhuanlan.zhihu.com/p/510539190 (accessed on 18 October 2025).

Appendix B

We present examples of stories generated by five multimodal large language models along with their corresponding prompts.

Table A1. The difference in generated stories by ChatGPT-4.0 using different prompts (the original text is in Chinese).

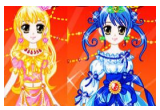
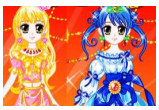
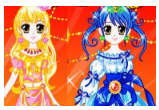
		
Prompt: According to the above image, please tell a children's fairy tale in English.	Prompt (English Translation): According to the above image, please tell a children's fairy tale in Chinese.	Prompt (English Translation): Please tell a relatively long Chinese children's fairy tale based on this picture, which includes several paragraphs.
Once upon a time, in the sparkling kingdom of Harmonia, music flowed through the air like magic. In this land lived two sisters, Melody and Harmony. Melody had golden hair and wore a dress as bright as the morning sun. Harmony had ocean-blue hair and wore a gown as deep and beautiful as the sea. The sisters had a very special gift—they could sing songs that made flowers bloom, animals dance, and even the stars twinkle brighter in the night sky. One day, the colors in Harmonia began to fade. The skies turned grey, and the rivers stopped singing. The kingdom's magic was disappearing! The queen summoned Melody and Harmony and said, "Only the Crystal Song, hidden in the Cave of Echoes, can restore our land." Bravely, the sisters set off on their journey. They traveled through forests that whispered secrets and crossed bridges made of rainbows. Finally, they reached the Cave of Echoes. Inside the cave, the Crystal Song was hidden inside a glowing gem. But a riddle guarded it: *"Only the sound of true love and harmony can unlock me."* Holding hands, the sisters sang from their hearts—a song of joy, courage, and sisterly love. The cave sparkled with light, and the crystal began to glow brighter than the sun. Suddenly, magic poured out of the cave, coloring the skies and waking the rivers. Harmonia was saved! From that day on, Melody and Harmony became known as the Magical Melody Sisters, protectors of music and color. And their songs? They still echo in the wind, wherever there's love and hope.	在遥远的彩光王国里，有一对神奇的双子公主。姐姐名叫露比雅，拥有如阳光般金灿灿的头发，性格温柔善良；妹妹名叫蓝蒂娅，头发像大空一样湛蓝，聪明勇敢。 彩光王国由七种颜色的魔法守护，红、橙、黄、绿、蓝、靛、紫。每一位公主出生时都会获得一种颜色的魔法。露比雅得到了"阳光橙"的治愈之力，可以治愈生病的植物和动物；蓝蒂娅则掌握了"星辰蓝"的智慧之力，能读懂古老的魔法文字。 有一天，王国的"彩虹之心"突然失去了光芒，整个天空变得灰暗。国王和王后十分着急，因为传说中只有"最纯洁的心灵"和"最坚定的智慧"才能重新唤醒彩虹之心。 于是，露比雅和蓝蒂娅踏上了旅途。她们穿越了沉睡森林、冰雪山谷，和迷雾平原，在旅途中，她们不断帮助别人——为一只生病的小鹿疗伤，照亮路的小精灵找到回家的路，还解开了一道古老的光之谜题。 终于，在彩虹山巅，她们找到了彩虹之心的神殿。露比雅用她纯净的心意唱出一段治愈之歌，蓝蒂娅则用智慧之光解开了最后的魔法封印，彩虹之心再次闪耀出七彩光芒，照亮了整个王国。 从此以后，彩光王国重新恢复了光明，而露比雅和蓝蒂娅也被称为"最闪耀的姐妹星"。她们继续在王国中传播爱与希望，守护每一个孩子的美梦。	(English Translation):In the distant Chromatic Kingdom, there lived a pair of magical twin princesses. The elder sister, named Rubia, had hair as golden as sunlight and a gentle, kind heart. The younger sister, Lantia, boasted hair as blue as the sky and was clever and brave. The Chromatic Kingdom was guarded by seven magical forces of color: red, orange, yellow, green, blue, indigo, and violet. Every princess was bestowed with a single color magic at birth. Rubia inherited the healing power of Sunlight Pink, which could cure sick plants and animals. Lantia, on the other hand, mastered the wisdom power of Stellar Blue, enabling her to decipher ancient magical texts. One day, the Rainbow Heart of the kingdom suddenly dimmed, and the entire sky turned gloomy. The king and queen fell into great anxiety, for an ancient legend foretold that only the "purest heart" and the "steadfastest wisdom" could reawaken the Rainbow Heart. Thus, Rubia and Lantia embarked on a journey. They traversed the Slumbering Forest, the Icy Snow Valley, and the Misty Plains. Along the way, they never hesitated to lend a helping hand—they healed a sick fawn, guided lost elves back home, and solved an age-old riddle of light. At long last, atop Rainbow Peak, they reached the Temple of the Rainbow Heart. Rubia sang a healing melody with her purest intention, while Lantia used her light of wisdom to break the final magical seal. The Rainbow Heart blazed forth with its seven-colored radiance once more, illuminating the entire kingdom. From then on, the Chromatic Kingdom regained its brightness. Rubia and Lantia were hailed as the Brightest Sister Stars. They continued to spread love and hope throughout the kingdom, safeguarding the sweet dreams of every child.	(English Translation): In a faraway kingdom, there lived two beautiful princesses: Princess Lucy with golden hair and Princess Lily with blue hair. They resided in a magnificent palace, surrounded by joy and singing every single day. Princess Lucy had long hair like threads of gold. She loved wearing a shimmering yellow gown and spent her days chasing butterflies and birds in the palace garden. Princess Lily, on the other hand, had hair as deep blue as the night sky. She always chose to wear a white dress adorned with blue gemstones and delighted in losing herself in the sea of books in the palace library. One day, a colossal dragon loomed over the kingdom. Its eyes blazed like two burning coals, and its scales glinted with a menacing light. Roaring thunderously, the dragon demanded that the kingdom offer it gold and silver treasures every month, threatening to destroy the land if its demand was refused. The king and queen were seized with great fear, yet they had no idea what to do. When Lucy and Lily heard the news, they resolved to do something to save their homeland. They knew that strength and courage did not lie merely in powerful weapons or brawny physiques, but in the hearts of those willing to stand up for love and justice. And so, they embarked on their adventure. The princesses first sought out a wise old witch, who told them that only the legendary Flower of Light could calm the dragon's wrath. The Flower of Light grew in a perilous place, protected by a formidable magic. Without a moment's hesitation, Lucy and Lily set off on their journey to find the flower. Along the way, they encountered all manner of difficulties and challenges, but their courage and wisdom helped them overcome each and every one. At long last, they found the Flower of Light, glowing with brilliant seven-colored radiance. They carefully carried the flower back to the kingdom and presented it to the fearsome dragon. The dragon was deeply moved by the flower's beauty. The anger and greed in its heart melted away. It pledged never to threaten the kingdom again and gave its own treasure to the land as a gift of peace. From then on, the dragon and humans lived in harmony. The king, the queen, and all the people were grateful to Lucy and Lily, who became heroes of the entire kingdom for their bravery and wisdom. Ever since, the story of Lucy and Lily has been sung throughout the land. Every child knows that true strength comes from a heart filled with love and courage. In the palace garden, the Flower of Light blooms forever, a reminder of the power of bravery and wisdom.

Table A2. The difference in generated stories by Qwen-VL-7B using different prompts (the original text is in Chinese).

	
Prompt (English Translation): According to the above image, please tell a children's fairy tale in Chinese.	Prompt (English Translation): Please tell a relatively long Chinese children's fairy tale based on this picture, which includes several paragraphs.
在一个美丽的国度里，有两个好朋友。她们是金发女孩和蓝头发女孩。她们非常可爱，也非常好相处。有一天，她们决定一起出去玩。金发女孩穿着她最喜欢的金色裙子，蓝头发女孩则穿着她最漂亮的蓝色裙子。她们走着走着，突然看到了一座美丽的城堡。城堡里传来了美妙的音乐声，让两个女孩非常好奇。于是，她们决定走进城堡一探究竟。城堡里的公主看到两个小女孩非常可爱，于是决定邀请她们一起跳舞。两个女孩非常开心，她们穿着漂亮的裙子，在城堡里翩翩起舞。这一天，两个女孩度过了一段非常美好的时光。她们不仅欣赏了美丽的风景，还结交了新的朋友。最后，她们依依不舍地离开了城堡，约定下次再来这里玩耍。	(English Translation):In a beautiful kingdom, there lived two good friends—a girl with golden hair and a girl with blue hair. They were utterly adorable and easy to get along with. One day, they decided to go out and play together. The golden-haired girl put on her favorite golden dress, while the blue-haired girl wore her prettiest blue dress. As they wandered along, they suddenly caught sight of a magnificent castle. Beautiful music drifted out from the castle, sparking the two girls' curiosity. So they made up their minds to step inside and find out what was going on. The princess in the castle thought the two little girls were very lovely, and she invited them to dance together. Overjoyed, the two girls twirled and danced gracefully in the castle, their beautiful dresses fluttering with every move. It was a truly wonderful day for the two girls. Not only did they enjoy the stunning scenery, but they also made a new friend. When it was time to leave, they parted with reluctance, making a promise to come back and play at the castle again.	从前有一个美丽的小村庄，里面住着许多可爱的小动物。有一天，村庄里的小公主艾米丽过生日，所有的动物们都来参加她的生日派对。艾米丽非常喜欢蓝色，所以她希望自己的生日派对是蓝色主题的。小动物们都非常支持艾米丽的想法，于是他们开始忙碌起来，准备蓝色的气球、蓝色的蛋糕和蓝色的装饰品。派对开始后，艾米丽非常开心，她和小动物们一起跳舞、一起唱歌，整个派对非常热闹。艾米丽收到了许多礼物，有蓝色的裙子、蓝色的鞋子和蓝色的发卡。她最喜欢的就是那个蓝色的发卡，它上面镶嵌着一颗红色的宝石，非常漂亮。艾米丽把发卡别在了自己的金色长发上，看起来更加美丽动人。生日派对结束后，艾米丽邀请小动物们到她的兴趣里做客，他们一起玩耍、一起吃饭，度过了一个非常快乐的下午。这个故事告诉我们，友谊和爱是最重要的，只要我们用心去对待身边的人和事，就会拥有一个美好的世界。	(English Translation):Once upon a time, there was a beautiful little village inhabited by many lovely little animals. One day, Princess Emily of the village was celebrating her birthday, and all the animals came to her birthday party. Emily loved the color blue dearly, so she wished for a blue-themed birthday party. The little animals all fully supported her idea and immediately got busy preparing blue balloons, a blue cake, and blue decorations. When the party started, Emily was overjoyed. She danced and sang with the little animals, and the whole party was filled with joy and excitement. Emily received plenty of gifts, including a blue dress, a pair of blue shoes, and a blue hairpin. Her favorite gift was the blue hairpin, which was inlaid with a bright red gemstone and looked absolutely stunning. Emily pinned the hairpin in her long golden hair, making her look even more beautiful and charming. After the birthday party, Emily invited all the little animals to her palace as guests. They played games and enjoyed a meal together, spending a truly delightful afternoon. This story tells us that friendship and love are the most precious things in the world. As long as we treat the people and things around us with sincerity, we will live in a wonderful world.

Table A3. The difference in generated stories by three LLMs using the same prompt (From left to right: VisCPM-10B, VCLA-7B, VisualGLM-6B. The original text is in Chinese.)

		
Prompt (English Translation): Please tell a relatively long Chinese children's fairy tale based on this picture, which includes several paragraphs.	Prompt (English Translation): Please tell a relatively long Chinese children's fairy tale based on this picture, which includes several paragraphs.	Prompt (English Translation): Please tell a relatively long Chinese children's fairy tale based on this picture, which includes several paragraphs.
在一个明亮的房间里有两个女孩，她们都穿着华丽的衣服。其中一个是蓝色头发，而另一个则是黄色头发，她们看起来像是洋娃娃一样，有着长长的金发和精致的五官，女孩们站得很近，似乎正享受着彼此的陪伴。 在她们身后的墙上挂着一幅画，上面是一只长着翅膀的猫头鹰。这只猫头鹰看上去就像真的一样，它展开了双翅，好像正在飞翔一样。这幅画为整个场景增添了一种奇思妙想和迷人的感觉。	(English Translation):In a bright room stood two girls, both dressed in gorgeous outfits. One had blue hair, while the other had golden hair. They looked just like dolls, with long flowing golden tresses and delicate features. The girls stood close to each other, seemingly savoring each other's company. On the wall behind them hung a painting depicting a winged owl. The owl looked remarkably lifelike, with its wings spread wide as if it were in mid-flight. This painting lent the entire scene a whimsical and enchanting vibe.	有一次，在一座美丽的城堡里，有一个名叫露丝的年轻女孩。她非常漂亮，而且非常聪明，很受朋友们的喜爱。有一天，露丝的父母告诉她，她们要去远方旅行，而她们会留下一个信封给她，里面装着她最喜欢的东西。露丝感到非常难过，因为她知道她的朋友们会离开她。当她的父母离去时，露丝打开了信封，里面是一个神奇的宝石。这个宝石可以把任何东西变成现实。露丝决定用它来创造一个美丽的花园，这样她就可以和她的朋友们一起度过时光。	(English Translation):Once upon a time, in a beautiful castle, there lived a young girl named Ruth. She was exceptionally beautiful and intelligent, and deeply beloved by her friends. One day, Ruth's parents told her that they were going on a long journey, and they would leave an envelope for her, which contained her favorite thing. Ruth felt very sad, for she thought her friends would leave her. When her parents departed, Ruth opened the envelope, and inside was a magical gemstone—one that could turn any wish into reality. Ruth decided to use it to create a beautiful garden, so that she could spend time there with her friends.	这张图片展示了两个穿着华丽礼服的卡通女孩站在红色背景前。其中一个身着粉色连衣裙和白色头巾，另一个则身穿蓝色裙子和紫色头巾。两人都戴着珠宝饰品，并装着蝴蝶结。她们都手持闪闪发光的戒指、项链和其他珠宝首饰。整个画面充满了梦幻般的氛围。	(English Translation):This picture shows two cartoon girls dressed in gorgeous gowns standing in front of a red background. One is wearing a pink dress with a white headscarf, while the other is in a blue dress with a purple headscarf. Both are adorned with jewelry ornaments and bows. They are each holding glittering rings, necklaces, and other pieces of jewelry. The entire scene is filled with a dreamlike atmosphere.

References

1. Zipes, J. *Fairy Tales and the Art of Subversion*, 2nd ed.; Routledge: London, UK, 2006.
2. Poole, C. The Importance of Fairy Tales for Children. 2022. Available online: https://thinkingwest.com/2022/04/25/importance-of-fairy-tales/ (accessed on 18 October 2025).
3. Wang, S. Problems and Reflections on Children's Literature Creation in the New Century. *Chin. Lit. Crit.* **2024**, *2*, 151–160.
4. Tan, X. *A Brief Discussion on Children's Literature*, 1st ed.; People's Posts and Telecommunications Press: Beijing, China, 2016.
5. VisikoKnox-Johnson, L. The Positive Impacts of Fairy Tales for Children. *Hohonu* **2016**, *14*, 77–81.
6. Winick, S. Einstein's Folklore | FolkLife Today. The Library of Congress. 2013. Available online: https://blogs.loc.gov/folklife/2013/12/einsteins-folklore/ (accessed on 18 October 2025).
7. OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2024**, arXiv:2303.08774. [CrossRef]
8. Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; Yang, D. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023.
9. Makridis, G.; Oikonomou, A.; Koukos, V. FairyLandAI: Personalized Fairy Tales utilizing ChatGPT and DALLE-3. *arXiv* **2024**, arXiv:2407.0946. [CrossRef]
10. Ye, L.; Jiang, J.; Chang, D.; Liu, P. Storypark: Leveraging Large Language Models to Enhance Children Story Learning Through Child-AI collaboration Storytelling. *arXiv* **2024**, arXiv:2405.06495. [CrossRef]
11. Xie, Z.; Cohn, T.; Lau, J.H. The Next Chapter: A Study of Large Language Models in Storytelling. In Proceedings of the 16th International Natural Language Generation Conference, Online, 13–15 September 2023; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 323–351.
12. Huang, H.; Tang, C.; Loakman, T.; Guerin, F.; Lin, C. Improving Chinese Story Generation via Awareness of Syntactic Dependencies and Semantics. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Taipei, Taiwan, 20–23 November 2022; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 178–185.
13. Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural language processing: State of the art, current trends and challenges. *Multimed. Tools Appl.* **2023**, *82*, 3713–3744. [CrossRef]
14. Xu, F.; Hao, Q.; Zong, Z.; Wang, J.; Zhang, Y.; Wang, J.; Lan, X.; Gong, J.; Ouyang, T.; Meng, F.; et al. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. *arXiv* **2025**, arXiv:2501.09686. [CrossRef]
15. Liang, X.; Wang, H.; Wang, Y.; Song, S.; Yang, J.; Niu, S.; Hu, J.; Liu, D.; Yao, S.; Xiong, F.; et al. Controllable Text Generation for Large Language Models: A Survey. *arXiv* **2024**, arXiv:2408.12599. [CrossRef]
16. Sun, J.; Tian, Y.; Zhou, W.; Xu, N.; Hu, Q.; Gupta, R.; Wieting, J.; Peng, N.; Ma, X. Evaluating Large Language Models on Controlled Generation Tasks. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 3155–3168. [CrossRef]

17. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels, Belgium, 1 November 2018; Linzen, T., Chrupała, G., Alishahi, A., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 353–355.
18. Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S.R. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Curran Associates Inc.: Red Hook, NY, USA, 2019; pp. 3266–3280.
19. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–45. [CrossRef]
20. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
21. Novikova, J.; Dušek, O.; Cercas Curry, A.; Rieser, V. Why We Need New Evaluation Metrics for NLG. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; Palmer, M., Hwa, R., Riedel, S., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 2241–2252.
22. Alabdulkarim, A.; Li, S.; Peng, X. Automatic Story Generation: Challenges and Attempts. In Proceedings of the Third Workshop on Narrative Understanding, Online, 11 June 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 72–83.
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6000–6010.
24. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 18 October 2025).
25. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. 2019. Available online: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (accessed on 18 October 2025).
26. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 7871–7880.
27. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume33, pp. 1877–1901.
28. Chen, H.; Shu, R.; Takamura, H.; Nakayama, H. GraphPlan: Story Generation by Planning with Event Graph. In Proceedings of the 14th International Conference on Natural Language Generation, Aberdeen, UK, 20–24 September 2021 ; Association for Computational Linguistic: Stroudsburg, PA, USA, 2021; pp. 377–386.
29. Yao, L.; Peng, N.; Weischedel, R.; Knight, K.; Zhao, D.; Yan, R. Plan-and-write: Towards better automatic storytelling. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu, HI, USA, January–1 February 2019; AAAI Press: Washington, DC, USA, 2019; pp. 7378–7385.
30. Walker, D.; Greenwood, C.; Hart, B.; Carta, J. Prediction of school outcomes based on early language production and socioeconomic factors. *Child Dev.* **1994**, *65*, 606–621. [CrossRef] [PubMed]
31. Valentini, M.; Weber, J.; Salcido, J.; Wright, T.; Colunga, E.; von der Wense, K. On the Automatic Generation and Simplification of Children's Stories. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023.
32. Bhandari, P.; Brennan, H. Trustworthiness of Children Stories Generated by Large Language Models. In Proceedings of the 16th International Natural Language Generation Conference, Prague, Czechia, 11–15 September 2023; Keet, C.M., Lee, H.Y., Zarrieß, S., Eds; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 352–361.
33. Nielsen, J. *Usability Engineering*; Academic Press: San Diego, CA, USA, 1993.
34. Driessen, G.; Smit, F.; Sleegers, P. Parental Involvement and Educational Achievement. *Br. Educ. Res. J.* **2005**, *31*, 509–532. [CrossRef]
35. Nastasiuk, A.; Courteau, E.; Thomson, J.; Deacon, S.H. Drawing attention to print or meaning: How parents read with their preschool-aged children on paper and on screens. *J. Res. Reading* **2024**, *47*, 412–428. [CrossRef]

36. Lin, C.; Šabanović, S.; Dombrowski, L.; Miller, A.D.; Brady, E.; MacDorman, K.F. Parental Acceptance of Children's Storytelling Robots: A Projection of the Uncanny Valley of AI. *Front. Robot. AI* **2021**, *8*, 579993. [CrossRef]
37. Sun, Y.; Chen, J.; Yao, B.; Liu, J.; Wang, D.; Ma, X.; Lu, Y.; Xu, Y.; He, L. Exploring Parent's Needs for Children-Centered AI to Support Preschoolers' Interactive Storytelling and Reading Activities. *Proc. ACM-Hum.-Comput. Interact.* **2024**, *8*, 1–25. [CrossRef]
38. OpenAI. ChatGPT. 2023. Available online: https://chat.openai.com.chat (accessed on 18 October 2025).
39. Hu, J.; Yao, Y.; Wang, C.; Wang, S.; Pan, Y.; Chen, Q.; Yu, T.; Wu, H.; Zhao, Y.; Zhang, H.; et al. Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages. In Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7 May 2024.
40. Cui, Y.; Yang, Z.; Yao, X. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. *arXiv* **2023**, arXiv:2304.08177. [CrossRef]
41. Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; Tang, J. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; Volume 1: Long Papers, pp. 320–335.
42. Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; Zhou, J. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv* **2023**, arXiv:2308.12966. [CrossRef]
43. Xu, S. Research on Teaching Strategies for Picture-Based Writing in Primary School Chinese Language Education. *Educ. Adv.* **2024**, *13*, 7218–7223. (In Chinese) [CrossRef]
44. Hui, Y.; Zhou, X.; Li, Y.; De, X.; Li, H.; Liu, X. Developmental Trends of Literacy Skills of Chinese Lower Graders: The Predicting Effects of Reading-related Cognitive Skills. *Psychol. Dev. Educ.* **2018**, *34*, 73–79. [CrossRef]
45. Riskin, J. *The Restless Clock: A History of the Centuries-Long Argument Over What Makes Living Things Tick*; University of Chicago Press: Chicago, IL, USA, 2016.
46. Castellano, G.; Vessio, G. Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. *Neural Comput. Appl.* **2021**, *33*, 12263–12282. [CrossRef]
47. Hernandez-Olivan, C.; Hernandez-Olivan, J.; Beltran, J.R. A Survey on Artificial Intelligence for Music Generation: Agents, Domains and Perspectives. *arXiv* **2022**, arXiv:2210.13944. [CrossRef]
48. Chakrabarty, T.; Padmakumar, V.; He, H. Help me write a Poem - Instruction Tuning as a Vehicle for Collaborative Poetry Writing. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 6848–6863. [CrossRef]
49. Mirowski, P.; Mathewson, K.W.; Pittman, J.; Evans, R. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 23–28 April 2023; CHI '23. [CrossRef]
50. Wu, W.; Wu, H.; Jiang, L.; Liu, X.; Zhao, H.; Zhang, M. From Role-Play to Drama-Interaction: An LLM Solution. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 3271–3290. [CrossRef]
51. Egg, A. An AI Learning Robot for Children, That Follows Along and Reads Whatever You Point at. 2016. Available online: https://www.toycloud.com (accessed on 18 October 2025).
52. Zhao, Z.; McEwen, R. Luka—Investigating the Interaction of Children and Their Home Reading Companion Robot: A Longitudinal Remote Study. In Proceedings of the Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, New York, NY, USA, 8–11 March 2021; HRI '21 Companion, pp. 141–143. [CrossRef]
53. Codi, M. An Interactive, AI-Enabled Smart Toy for Kids. 2021. Available online: https://www.pillarlearning.com/ (accessed on 18 October 2025).
54. Dietz, G.; Le, J.K.; Tamer, N.; Han, J.; Gweon, H.; Murnane, E.L.; Landay, J.A. StoryCoder: Teaching Computational Thinking Concepts Through Storytelling in a Voice-Guided App for Children. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 8–13 May 2021; CHI '21. [CrossRef]
55. Zhang, Z.; Xu, Y.; Wang, Y.; Yao, B.; Ritchie, D.; Wu, T.; Yu, M.; Wang, D.; Li, T.J.J. StoryBuddy: A Human-AI Collaborative Chatbot for Parent-Child Interactive Storytelling with Flexible Parental Involvement. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 29 April–5 May 2022; CHI '22. [CrossRef]
56. Yi, R. *Text and Image Studies on AIGC Productivity in the Age of Humanity 3.0*; China Social Sciences Press: Beijing, China, 2024.
57. Liu, J.; Gu, H.; Zheng, T.; Xiang, L.; Wu, H.; Fu, J.; He, Z. Dynamic Generation of Personalities with Large Language Models. *arXiv* **2024**, arXiv:2404.07084. [CrossRef]
58. Porter, B.; Machery, E. AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Sci. Rep.* **2024**, *14*, 26133. [CrossRef]
59. Liu, Z.; Li, Y.; Cao, Q.; Chen, J.; Yang, T.; Wu, Z.; Hale, J.; Gibbs, J.; Rasheed, K.; Liu, N.; et al. Transformation vs Tradition: Artificial General Intelligence (AGI) for Arts and Humanities. *arXiv* **2023**, arXiv:2310.19626. [CrossRef]

60. Rahman, A.; Mahir, S.H.; Tashrif, M.T.A.; Aishi, A.A.; Karim, M.A.; Kundu, D.; Debnath, T.; Moududi, M.A.A.; Eidmum, M.Z.A. Comparative Analysis Based on DeepSeek, ChatGPT, and Google Gemini: Features, Techniques, Performance, Future Prospects. *arXiv* **2025**, arXiv:2503.04783. [CrossRef]
61. Chen, X.; Hu, Z.; Wang, C. Empowering education development through AIGC: A systematic literature review. *Educ. Inf. Technol.s* **2024**, *29*, 17485–17537. [CrossRef]
62. Abdelghani, R.; Wang, Y.H.; Yuan, X.; Wang, T.; Lucas, P.; Sauzéon, H.; Oudeyer, P.Y. GPT-3-Driven Pedagogical Agents to Train Children's Curious Question-Asking Skills. *Int. J. Artif. Intell. Educ.n* **2023**, *34*, 483–518. [CrossRef]
63. Kerneža, M. Fundamental and basic cognitive skills required for teachers to effectively use chatbots in education. In *Science And Technology Education: New Developments And Innovations*; Scientia Socialis, UAB: Siauliai, Lithuani, 2023; pp. 99–110. [CrossRef]
64. Holmes, W.; Kay, J. AI in Education. Coming of Age? The Community Voice. In Proceedings of the Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky, Tokyo, Japan, 3–7 July 2023; pp. 85–90. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.