Can Large Language Models (LLMs) Describe Pictures Like Children? A Comparative Corpus Study.

Anonymous ACL submission

Abstract

Large language models (LLMs) are applied to diverse contexts of our lives, including the implementation in child education. Here, we evaluate the ability of an LLM to generate child-like language by comparing an LLM-based corpus to the Litkey Corpus, a collection of German children's writings based on picture stories. We generated a parallel LLM-based corpus using identical visual prompts and conducted a comparative analysis across word frequency distributions, lexical richness, and semantic representations. This study aims to explore if and how children and LLMs differ in psycholinguistic aspects of the text to evaluate the potential influences of LLM text on child development. The results show that, while the LLM-based texts are longer, the vocabulary is less rich, has more letters, and misses words in medium- and low-frequency ranges (i.e., uses primarily words that often occur). Additionally, vector space analysis using semantic word embeddings reveals a low semantic similarity, highlighting differences between the two corpora on the level of corpus semantics. These findings contribute to our understanding of LLM-generated language and its limitations in modeling child language, with implications for LLM usage in psycholinguistics and educational applications.

1 Introduction

Large Language Models (LLMs) have impacted fields like corpus linguistics, psycholinguistics, and natural language processing (NLP) by providing new ways to generate text in response to prompts (Brown, 2020; Bommasani et al., 2021; Devlin, 2018; Vaswani et al., 2017). High user-friendly usage of LLMs increased the use of such models in research and in applied settings (e.g., as implementations in chatbots; (Dam et al., 2024; Brown, 2020)). For example, recent developments used LLMs to train children's creative writing skills (Elgarf et al., 2024). The finding is that one can boost child creative writings based on LLM output as recent investigations indicated that LLM-generated text lacks lexical richness (Liu and Fourtassi, 2024; Schepens et al., 2023). Here, we extend this evidence based on the latest state-of-the-art models that offer image prompt capabilities, generating text from both text and visual prompts (Tsimpoukelli et al., 2021; Alayrac et al., 2022). These LLMs allow us to simulate a unique yet child-specific corpus that collects children's writings in response to a set of picture stories.

The study of child language and first language acquisition is foundational in linguistics, providing insights into cognitive development and the mechanisms of language learning (Tomasello, 2005; Clark and Casillas, 2015). During early development, children undergo significant linguistic and interactive development, making this period critical for understanding human language acquisition. However, the representation of child language in computational and NLP research remains limited due to challenges such as ethical restrictions, difficulty in data collection, and limited corpora availability (MacWhinney, 2000; Casillas et al., 2017). LLMs primarily rely on large-scale datasets dominated by adult and high-resource language data (Luo et al., 2023). Child language remains marginal in the models' training data, especially in underrepresented languages (but see, e.g., (Warstadt et al., 2023; Thoma et al., 2023)). As a result, there is limited research on whether LLMs can efficiently generate child-like text that mimics the linguistic patterns and conceptual structures of children's writing (but see Liu and Fourtassi (2024); Schepens et al. (2023)).

To address this gap, this study compares an LLM-generated corpus to the Litkey Corpus of German children's texts (Laarmann-Quante et al., 2019b). This comparison offers insights into how well LLMs simulate child language and provides a framework for understanding the implications of using LLM-generated corpora in linguistic and NLP contexts. Moreover, in psycholinguistic research, corpora generated by LLMs for underrepresented languages and groups have the potential to expand current research beyond the predominantly studied populations and high-resource languages (Blasi et al., 2022; Henrich et al., 2010; Gagl et al., 2022).

Our study investigates if LLMs and children generate similar descriptions of the exact same picture stories used in the Litkey Corpus (stories can be downloaded from: Litkey Corpus). A comparative analysis examines both corpora in terms of word frequency (Schepens et al., 2023; Brysbaert et al., 2011; Schroeder et al., 2015), lexical diversity (Schepens et al., 2023; Baayen and Baayen, 2001; Keuleers et al., 2015), and vector-based semantic similarity with word embeddings (Babić et al., 2020; Bojanowski et al., 2017; Günther et al., 2019). The study aims to determine whether LLMs can replicate the linguistic features of child language and assess potential risks for application in educational settings and future challenges for model development. By analyzing these patterns, we expect to uncover both similarities and key differences that provide insight into the capabilities and limitations of LLMs in modeling child-like language.

2 Methods

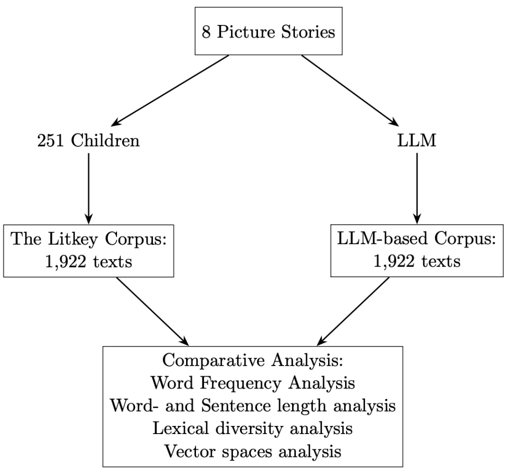

Figure 1: Methodology overview: Comparative analysis of the Litkey Corpus created from children's descriptions and the LLM-based corpus generated from the same picture stories.

Figure 1 provides an overview of the current study. Using identical visual stimulation (eight picture stories, see Litkey Corpus), 251 children (i.e., the Litkey Corpus (Laarmann-Quante et al., 2019b)) and one LLM produced 1,922 texts (i.e., GPT-4 (Achiam et al., 2023)). In addition to the image prompts, the LLM was instructed to generate text simulating the writing of children at the average age from the Litkey corpus. Thus, we generated two corpora that can be compared on the basis of well-established psycho-linguistic concepts: (i) word frequency (i.e., how often a word occurs in the corpus), word length (i.e., how many letters do words have), sentence length (i.e., number of words per sentence), lexical diversity (i.e., as measured by log-TTR), and semantic similarity (i.e., using semantic vector space representations from word embeddings).

2.1 The Litkey Corpus

The Litkey Corpus is a richly annotated collection of German children's texts designed to highlight the later stages of orthographic and literacy development (Laarmann-Quante et al., 2019b). The corpus enables a comprehensive analysis of language acquisition across diverse learner backgrounds (Laarmann-Quante et al., 2019a).

The corpus consists of texts written by 251 German primary school students, collected between 2010 and 2012, with an average age of the participants of 9.6 years (grades 2–4). It is based on eight picture stories comprising six textless images featuring three recurring characters. The corpus consists of 1,922 texts with 212,505 tokens and 6,364 types. We focus our analysis on all orthographic corrected versions of the texts with at least 15 readable words (Laarmann-Quante et al., 2019b). The collection process involved a 10-minute discussion about the pictures between the children and 30 minutes of writing. No story details were provided except for the main characters' names (Lars, Lea, Dodo). Thus, we assume that the Litkey Corpus is a robust resource for our planned comparison study.

2.2 Large Language Model

As LLM, we use *GPT-4V for vision*, a transformer architecture that supports text and image prompts (Achiam et al., 2023). A common problem with high-end LLMs is that key architectural details and model weights are not disclosed. Still, at the time of the study (March 2024), no much preferred open-source alternative was available. The model was accessed via the GPT-4V API, in Python. The cost of using the model was approximately $0.03 per 1,000 input tokens and $0.06 per 1,000 output tokens, with a token limit of 8,192 per API call. Visual inputs, encoded as Base64, contribute to this token limit. Larger or more complex images consume more tokens, potentially reducing the length of generated text outputs. It was manageable due to the naturally short texts in this study, but this limit can pose challenges for research on longer texts. Technical issues, such as a recurring `KeyError:'choices'` related to token allocation for visual inputs, were mitigated by setting the `max_tokens` parameter to 2,000. However, occasional errors persisted due to API refinements.

2.3 Prompt Engineering and the LLM-based Corpus

Prompt engineering is critical and allows us to generate text that should replicate child-specific text (e.g., see (Schepens et al., 2023)). Here, we prompt the LLM to generate child-like descriptions of the Litkey picture stories in German. In addition, to align outputs with the original data even more, we included the average age of the children as a parameter (9.6 years). After we tested different prompt structures to ensure the model generated language replicating written descriptions of children rather than simulating or "imagining" how a child might write, we ended up with the following structure that combined a visual input (i.e., picture stories) encoded in Base64 format and the following text:

"`Du bist ein {age}-jähriges Kind. Wie würdest du dieses Bild beschreiben?`" ("You are a {age}-year-old child. How would you describe this picture?").

For these prompts, we adjusted the following parameters: the `max_tokens` parameter was set to 2,000 to allow for sufficient text generation, while the `temperature` was set to 0.7 to encourage varied but coherent outputs. All other parameters (e.g., `presence_penalty` or `frequency_penalty` with the default setting of 0) were not manipulated explicitly. The total cost of generating these texts, including input and output tokens, was approximately $33. The final LLM-based corpus contained 1,922 texts with 363,867 tokens (averaging 189 tokens per text) and 3,855 types (see Table 1).

2.4 Data Preprocessing and Comparative Analysis

Both corpora were tokenized using the NLTK text mining library in Python ((Bird et al., 2009), with the language specified as `word_tokenize(text, language="german")`). Minimal preprocessing was applied to preserve the raw characteristics of the child- and LLM-generated texts. Lowercasing was avoided to retain the distinction between capitalized nouns (*Lernen*) and verbs (*lernen*), maintaining the structural integrity and better comparability of the German corpora (e.g., see (Schepens et al., 2023)).

The comparative analysis covered word frequency, lexical diversity, word and sentence length, and semantic vector space comparisons. The word frequency overview included the most frequent words with their counts in each corpus and those shared between them, including a list of the 10 most frequent words with >10 characters (see Table 3). Lexical richness was measured using the log type-token ratio (log TTR or Herdan's C) (Tweedie and Baayen, 1998; Herdan, 1960), where the log transformation accounts for text length:

$$\text{log-TTR} = \frac{\log(\text{Types})}{\log(\text{Tokens})} \quad (1)$$

Word frequency distributions of both corpora were compared to Zipf's Law to confirm natural language patterns, where a few high-frequency words are most common while many words appear infrequently. Additionally, we counted the letter length of words and the word length of sentences.

2.5 Vector-Based Semantic Analysis with Word Embeddings

The corpus comparison on the level of semantics relies on how words are distributed in a multidimensional vector space. This involves generating word embeddings, a technique rooted in distributional semantics, where semantically similar words are closer to each other (Elman, 1990; Firth, 1957; Emerson, 2020). The analysis utilized vector-based semantic analysis based on word embedding models trained using neural networks (e.g., GloVe (Pennington et al., 2014), Word2Vec (Mikolov, 2013)). Here, we use the fastText model (Bojanowski et al., 2017) due to its ability to capture subword information, making it particularly effective for morphologically rich languages like German and smaller corpora (Bojanowski et al., 2017). Subword information is also valuable for analysis that include child language because, even in the presence of spelling or orthographic errors, fastText can still accurately capture the word's meaning (Grave et al., 2018).

We used the preprocessed data from both corpora to train a fastText model instead of using pretrained word embeddings. Training a model on a specific corpus provides representations tailored to that corpus, capturing the nuances of the text. Thus, one set of embeddings reflected the unique characteristics of the Litkey Corpus and a second set of the LLM-based corpus. After training, all words from both corpora were converted into vectors (i.e., word embeddings) by the trained fastText model. The semantic similarity between words in both corpora was then assessed using cosine similarity, a widely used method to measure semantic similarity between vectors. Cosine similarity calculates the cosine of the angle between two vectors, with scores ranging from -1 (completely opposite) to 1 (identical). To compare the corpora, we calculated the cosine similarity between all shared words within each corpus. This cosine similarity with each corpus allows us to correlate the semantic relations between words across corpora since each corpus-specific vector space is arbitrary and cannot be compared.

3 Results

Corpus:	**Litkey**	**LLM-based**
Total texts	1,922	1,922
Total tokens	212,505	363,867
Avg. tokens/text	111	189
Total types	6,364	3,855
log-TTR	0.71	0.64

Table 1: Number of texts, total tokens, average tokens per text, total types, and lexical richness measured by log-TTR from the Litkey and the LLM-based corpora.

3.1 Frequency Distributions and Lexical Richness

Both corpora consist of the same number of texts. Still, the LLM-based corpus contains significantly more tokens than the Litkey Corpus, and the Litkey Corpus has a higher number of unique types, indicating greater lexical richness due to more unique words (see Table 1). The log-TTR is higher for the Litkey compared to the LLM-based corpus (cp. 0.71 vs. 0.64), indicating a higher lexical richness for the writings from children. Still, we found longer texts in the LLM compared to the Litkey corpus (mean number of tokens: 189 vs. 111 per text).

Litkey	**Litkey > 10**	**LLM-based**	**LLM-based > 10**
und	Staubsauger	und	Staubsauger
Lea	verschwunden	Bild	telefoniert
Dodo	erschrocken	der	Kuscheltier
Lars	Staubsaugerbeutel	das	erschrocken
ist	Fensterbank	ist	Bildergeschichte
sie	weggelaufen	Im	wahrscheinlich
hat	Hundefutter	ein	Staubsaugerbeutel
der	Telefonnummer	Hund	Klassenzimmer
die	Steckbriefe	sieht	wiederfindet
den	mitgebracht	aus	Comic Geschichte

Table 2: Comparison of the 10 most frequent words overall and words with >10 characters in the Litkey Corpus and the LLM-based corpus.

In Table 2, the two left columns display the 10 most common words from both corpora. Words such as *und* (and), *ist* (is), *der* (masculine article), and *die* (feminine/plural article) appear frequently in both corpora, reflecting their syntactic importance in German. The analysis of complex words (two right columns in Table 2) reveals further differences. In the Litkey Corpus, words such as *Staubsauger* (vacuum cleaner), *erschrocken* (scared), and *Telefonnummer* (phone number) are more specific and complex and indicate that children provided more detailed descriptions. In contrast, in the LLM-based corpus, more generic words like *Kuscheltier* (stuffed animal) and *Comic Geschichte* (comic story) are prevalent.

Table 3 compares the 20 most frequent words shared between the two corpora. Despite these shared words, differences in the use of content words are apparent. A detailed comparison showed that approximately 24.4% of the words are shared between the two corpora. While common function words dominate both corpora, the Litkey Corpus exhibits a wider variety of specific nouns, likely due to the children's imaginative and observational input. In the Litkey Corpus, the high frequency of names like *Dodo*, *Lea*, and *Lars* is notable, as these characters were introduced to the children before writing. Interestingly, the LLM-based corpus includes all three character names, even though these names were not provided in the prompt. However, their frequency is significantly lower compared to

Word	Both corpora	Count LLM	Count Litkey
und	24,036	14,205	9,831
der	13,240	10,189	3,051
Bild	12,683	12,635	48
ist	11,298	7,786	3,512
das	10,450	8,632	1,818
Lea	9,058	7	9,051
Dodo	8,812	221	8,591
ein	8,708	6,552	2,156
sie	7,574	4,082	3,492
Hund	7,524	6,549	975
Im	7,509	7,490	19
die	7,471	4,496	2,975
Lars	7,413	7	7,406
aus	6,918	6,090	828
sieht	6,517	6,242	275
dem	6,274	5,340	934
auf	5,995	3,729	2,266
den	5,593	3,202	2,391
hat	5,272	2,115	3,157
mit	4,559	3,097	1,462

Table 3: The 20 most frequent words shared in the Litkey Corpus and the LLM-based corpus.

the Litkey Corpus, possibly because the model recognized these names from a reference in the picture stories (the pictures are available at: Litkey Corpus). Two words dominate the LLM corpus: *Bild* (picture) and *im* (in). The difference in frequencies lies in the fact that model starts almost every description with *Im ersten Bild...* (In the first picture...), *Im zweiten Bild...* (In the second picture...), a phrasing rarely used by children.

3.2 Detailed Comparison to the Litkey Corpus

To analyze the relationship between the two corpora, the correlation between word frequencies was computed, followed by normalization and log transformation of the word frequencies. This procedure is a long-standing standard allowing optimal investigation of word frequency measures (e.g., see (Brysbaert and New, 2009; Schepens et al., 2023)) that reduces the dominant effect of a low number of high-frequency words (e.g., *und* (and), *ist* (he/she/it is), *er* (he)). Thus, it focuses strongly on the overall distribution of medium- and lower-frequency words. A Laplace smoothing transformation was applied to ensure all word counts were at least 1, avoiding issues with computing the logarithm of zero due to words present in one corpus but absent in the other:

$$f_{\text{log}} = \log(f_{\text{word}} + 1) \qquad (2)$$

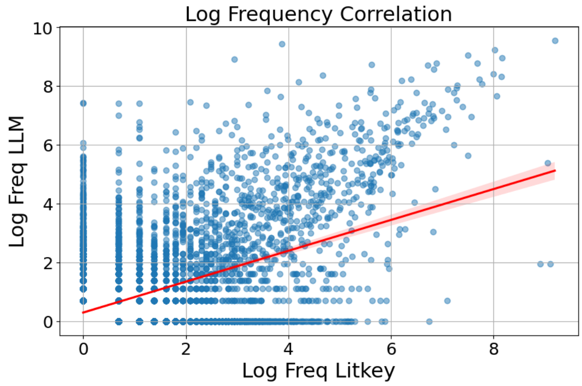

Figure 2: Log-normalized correlation between word frequencies in the LLM-based corpus and the Litkey Corpus showing a correlation of r = .47.

where f_{word} is the raw frequency of a word in the corpus. To obtain the f_{log}, we add 1 to the raw frequency (i.e., smoothing) and implement a logarithmic transformation to normalize the distribution and reduce the influence of high-frequency words.

After normalizing the word frequencies for corpus size, the correlation is 0.51, demonstrating that the relationship between the two corpora remains consistent. However, the log-normalized correlation (Figure 2) was slightly lower at 0.47, reflecting the model's difficulty in mimicking the distribution of less frequent words and suggests that, while the LLM can replicate common words used by children, it struggles with more unique or context-specific words that children use.

3.3 Alignment with the Zipf's Law

Figure 3 illustrates the comparison between the Litkey Corpus and the LLM-based corpus in terms of Zipf's law. Both corpora exhibit the expected inverse relationship between word rank and frequency, consistent with Zipf's law, which means that they contain a few high-frequency words and many low-frequency words, reflecting typical natural language patterns.

The X-axis in Figure 3 represents word rank (sorted by frequency) on a logarithmic scale, where lower ranks correspond to more frequent words. The Y-axis represents word frequency, also on a logarithmic scale, with higher values indicating more frequent words.

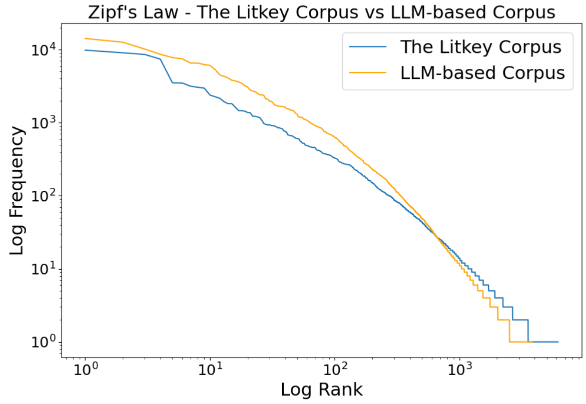

Figure 3: Zipf's law comparison for the Litkey Corpus and the LLM-based corpus, showing the relationship between word rank and frequency on a logarithmic scale.

The LLM-based corpus shows higher frequencies for the most frequent words (left side of the graph) compared to the Litkey corpus. This suggests that the LLM relies on a smaller set of high-frequency words that have been present more often. In contrast, the Litkey Corpus exhibits a richer distribution of less frequent words, as indicated by its curve overtaking that of the LLM-based corpus in the mid- and low-frequency ranges. This highlights the greater lexical diversity and variety of rare words in the children's texts compared to the LLM-generated texts.

3.4 Word and Sentence Length Distribution

Figure 4 compares the distributions of word and sentence lengths in the Litkey and the LLM-based corpora.

Figure 4 (left panel) shows that children's writing predominantly consists of short words (mean word length in the Litkey = 3.96). The LLM-based corpus also peaks at four characters but shows a slightly higher word length overall (mean word length in LLM-based corpus = 4.08), reflecting that the model generates texts with longer words. Figure 4 (right panel) shows the distribution of sentence length, indicating that the LLM-based corpus has longer sentences (mean sentence length in LLM-based corpus = 16.76 vs. mean sentence length in the Litkey = 12.59). At the same time, the Litkey corpus has a lower variability (standard deviation word length in LLM-based corpus = 2.41 vs. standard deviation word length Litkey = 2.06). Sentences exceeding 50 words are rare in the Litkey Corpus, but some outliers exist. These texts often lacked punctuation entirely, resulting in unusually long sentences. As only spelling errors were corrected, punctuation errors remained unchanged. 29 outliers exceeded 100 words, with the longest sentence having 278 tokens, and were excluded from the graph.

3.5 Vector Space Analysis

After training fastText-based word embeddings on both corpora, we calculated the cosine similarity between all shared words within both corpora. Based on a similar representation (e.g., Edelman (1998); Kriegeskorte et al. (2008)), we have a comparison based on an abstract representation that is independent of the original arbitrary embeddings. When we correlate the cosine similarities from both corpora, we estimate the semantic similarity between as r = .10 (see Fig. 5).

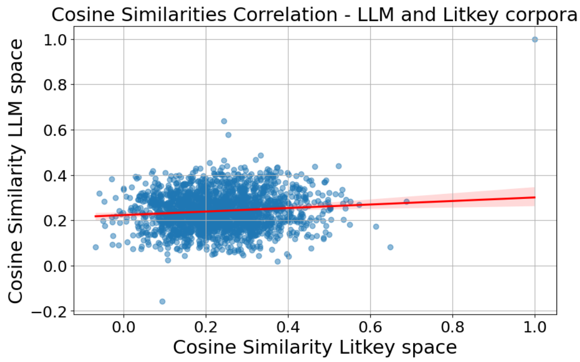

Figure 5: Cosine similarity between Litkey and LLM-based corpora (bootstrapped).

4 Discussion

This study compares written text from German children describing picture stories to Large Language Model-generated text prompted with the same images based on multiple established corpus- and psycho-linguistic measures. Our analysis shows that significant differences exist while LLMs can partially replicate children's writing. We found central differences in the characteristics of words, sentences, texts, and semantics. For words, the LLM-based corpus involves orthographically correct, longer words and has lower lexical richness. Obviously, LLMs are much better at producing words with correct spelling compared to children. Low lexical richness indicates that LLMs generate repetitive and less detailed vocabulary than children. Still, the LLM-generated text followed Zipf's law in general. However, detailed inspections indicated a higher number of high-frequency words (more words that occur often) and a lower number of mid- to low-frequency words (less rare words).

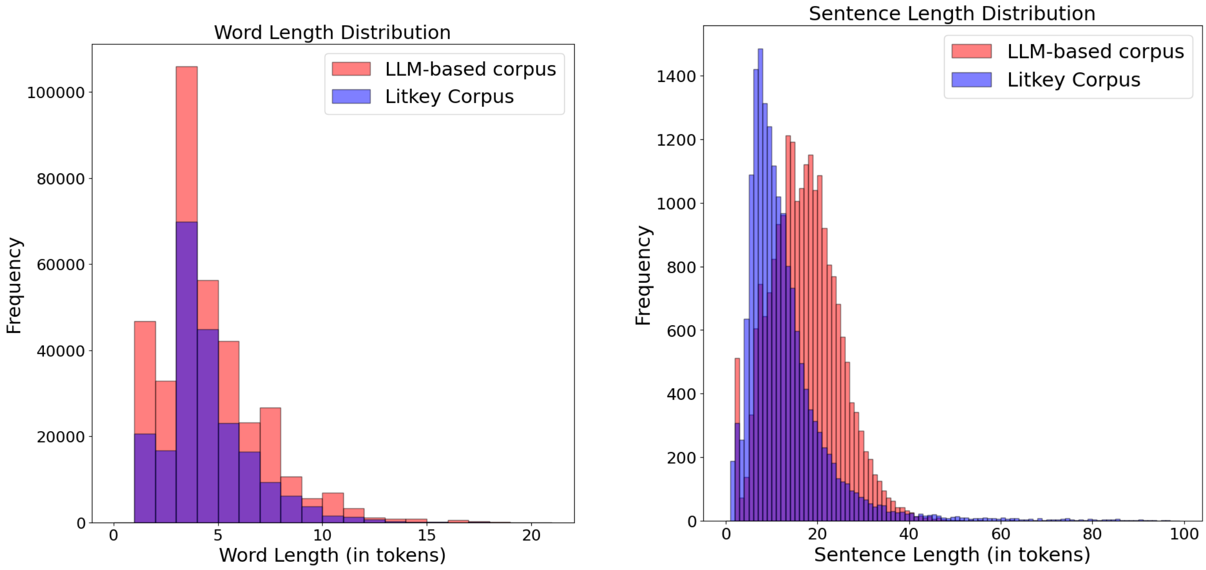

Figure 4: Distributions of word length (in the number of letters) and sentence length (in the number of words) of the Litkey and the LLM-based corpora.

When correlated, the word frequency measured from either corpus was respectable at an r of .47. On the level of sentences and texts, we found that the LLM generated longer sentences and texts than children. For semantics, we found that word co-occurrence-based semantic vector spaces differed drastically between the two corpora, indicating substantial differences in the semantic structure transported by LLMs and children. Thus, despite generating more prolonged, syntactically more complex texts (i.e., indexed by sentence length) replicating general lexical patterns (i.e., Zipf's law), the model struggled to replicate the diverse, context-rich language of children's storytelling, especially on the level of vocabulary and semantics.

A closer qualitative investigation of the LLM-generated texts showed that the general plot was captured (i.e., event sequence depicted in the pictures). Still, the LLM often misidentified key objects and characters, leading to inaccuracies and hallucinations. Therefore, we interpreted these findings as model hallucinations (e.g., fabricated text on posters or misidentifying characters or objects; see also Weidinger et al. (2022); Ji et al. (2023); Bender et al. (2021)). These errors stem from the model's tendency to process images in isolation rather than integrating context across the sequence. The frequent use of framing phrases like *Auf dem Bild sehe ich...* (In the picture, I can see...) contrasted with children's more spontaneous, narrative-focused descriptions. While applicable in structured educational contexts, this reliance on introductory phrases limited the model's ability to produce natural, flowing narratives.

This investigation is broader than previous ones but replicates central aspects. Lower lexical richness was previously found when human face-to-face interactions, child-book-based corpora, and direct comparison of general adult writings were compared to LLM-generated text (Schepens et al., 2023; Liu and Fourtassi, 2024; Guo et al., 2023). This evidence could be the result of a regression to the mean phenomenon, suggesting that the model training results in less lexical-rich outputs as a potential main goal of the LLM technology is to be comprehended by most people. In a first experimental approach, Schepens et al. (2023) experimented with increasing the temperature parameter, finding a slight increase in lexical richness. This indicates that parameter tuning of the models could be the first step towards better-representing child language in LLM models. Until such a set of parameters is not present and evaluated, or alternative evaluated approaches involving fine-tuning or the implementation of child-specific models, it is not recommended to use LLM agents in a child education context, as such could bias children's language towards a less rich vocabulary, potentially impoverishing development.

One goal of children's books is to increase vocabulary. A central finding in vocabulary research is that at the onset of literacy acquisition, the vocabulary increases drastically (Verhoeven et al., 2011; Song et al., 2015). Suppose we want children to be up to the task of recognizing language at a high quality in adulthood. In that case, we should present less rich, highly understandable text in only a few contexts related to instruction-type texts (i.e., how to implement a math problem or how to use a fire extinguisher; i.e., see Schepens et al. (2023) for a discussion) but not when it comes to educational language content or creative writing (e.g., (Elgarf et al., 2024)). We are in line with Liu and Fourtassi (2024) that, at present, we first need to design a set of benchmarks on which we can evaluate LLM-generated text before we use these models in educational or child-directed contexts.

One aspect indicated by Guo et al. (2023) is that the use of LLM-generated text reduces lexical richness. Data shortage was proposed to be accounted for by training on self-generated texts (Wang et al., 2022). For children, a natural shortage of child-produced text for training could be one reason for the low lexical richness and the poor performance on psycholinguistics measures for simulating parent-child interactions (Liu and Fourtassi, 2024), adult-written children books (Schepens et al., 2023) and children writings, as shown here.

5 Conclusions

We find that LLM-generated descriptions of picture stories, based on prompts that should result in child-like texts, are capable of generating orthographically correct but psycholinguistically very different text. We found fundamental differences between text produced by children and LLMs in response to the same picture stories on the level of words (e.g., lexical richness), sentences (i.e., number of words), texts (i.e., text length), and semantics (i.e., semantic vector spaces similarity). While large language models show promise in replicating certain linguistic features of children's language, they lack the full range of expressive and descriptive abilities. These findings highlight the need to continue exploring LLMs' capabilities and limitations to provide beneficial benchmarks for future developments.

6 Limitations

A significant limitation of this study lies in the lack of transparency of the used proprietary LLM (GPT-4v). Thus, we cannot evaluate details about training data, model parameters, and architectural parameters that could be relevant to the outcome of the study. With the spike in available open-access models, other open LLM models with image input available can be used here. We started to test the DeepSeek model (Bi et al., 2024) which, at the early stage of testing, seems to be a promising alternative in replicating child language (for an example description of one of the picture stories, see: A). In addition, closed-source models have opaque updates and modifications, drastically affecting performance (e.g., see Maslej et al. (2024); Chen et al. (2023); Schepens et al. (2023). Without this information, proprietary model outputs' replicability and reproducibility rates will stay low. So, to resolve these issues, we need more multimodal LLMs that are open to investigation in detail, and that can be frozen in a particular version.

Another limitation is that we have not covered syntactic (Liu and Fourtassi, 2024) and readability scores yet (Crossley et al., 2023). Broadening the benchmarking measures will allow a more holistic inspection of the generated LLM-generated text. Further, German as the main language might can pose a challenge here in relation to the amount of training data available for underrepresented languages, a similar issue to the lack of training resources for children's language development. Nonetheless the findings from a conversational corpus in English showed similar differences (Liu and Fourtassi, 2024) indicate that this issue is likely language in general. Still, future research should include a more holistic set of measures that includes all major aspects of language and, at the same time, investigate these issues across languages within one study.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

R Harald Baayen and R Harald Baayen. 2001. Word frequencies. *Word Frequency Distributions*, pages 1–38.

Karlo Babić, Francesco Guerra, Sanda Martinčić-Ipšić, and Ana Meštrović. 2020. A comparison of approaches for measuring the semantic similarity of short texts based on word embeddings. *Journal of information and organizational sciences*, 44(2):231–246.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Damián E Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 2022. Over-reliance on english hinders cognitive science. *Trends in cognitive sciences*, 26(12):1153–1170.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect. *Experimental psychology*.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Marisa Casillas, Elika Bergelson, Anne S Warlaumont, Alejandrina Cristia, Melanie Soderstrom, Mark VanDam, and Han Sloetjes. 2017. A new workflow for semi-automatized annotations: Tests with long-form naturalistic recordings of childrens language environments. In *Interspeech 2017*, pages 2098–2102.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Eve V Clark and Marisa Casillas. 2015. First language acquisition. In *The Routledge handbook of linguistics*, pages 311–328. Routledge.

Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2023. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55(2):491–507.

Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shimon Edelman. 1998. Representation is representation of similarities. *Behavioral and brain sciences*, 21(4):449–467.

Maha Elgarf, Hanan Salam, and Christopher Peters. 2024. Fostering children's creativity through llm-driven storytelling with a social robot. *Frontiers in Robotics and AI*, 11:1457429.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Guy Emerson. 2020. What are the goals of distributional semantics? *arXiv preprint arXiv:2005.02982*.

JR Firth. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis, Special Volume/Blackwell*.

Benjamin Gagl, Klara Gregorova, Julius Golch, Stefan Hawelka, Jona Sassenhagen, Alessandro Tavano, David Poeppel, and Christian J. Fiebach. 2022. Eye movements during text reading align with the rate of speech production. *Nature Human Behaviour*, 6:429–442.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. The curious decline of linguistic diversity: Training language models on synthetic text. *arXiv preprint arXiv:2311.09807*.

Fritz Günther, Luca Rinaldi, and Marco Marelli. 2019. Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, 14(6):1006–1033.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Gustav Herdan. 1960. Type-token mathematics: A textbook of mathematical linguistics. *(No Title)*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Emmanuel Keuleers, Michaël Stevens, Paweł Mandera, and Marc Brysbaert. 2015. Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, 68(8):1665–1692.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249.

Ronja Laarmann-Quante, Stefanie Dipper, and Eva Belke. 2019a. The making of the litkey corpus, a richly annotated longitudinal corpus of german texts written by primary school children. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 43–55.

Ronja Laarmann-Quante, Katrin Ortmann, Anna Ehlert, Simon Masloch, Doreen Scholz, Eva Belke, and Stefanie Dipper. 2019b. The litkey corpus: A richly annotated longitudinal corpus of german texts written by primary school children. *Behavior Research Methods*, 51:1889–1918.

Jing Liu and Abdellah Fourtassi. 2024. Benchmarking llms for mimicking child-caregiver language in interaction. *arXiv preprint arXiv:2412.09318*.

Queenie Luo, Michael J Puett, and Michael D Smith. 2023. A" perspectival" mirror of the elephant: Investigating language bias on google, chatgpt, youtube, and wikipedia. *arXiv preprint arXiv:2303.16281*.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. 2024. Artificial intelligence index report 2024. *Preprint*, arXiv:2405.19522.

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Job Schepens, Hanna Woloszyn, Nicole Marx, and Benjamin Gagl. 2023. Can we utilize large language models (llms) to generate useful linguistic corpora? a case study of the word frequency effect in young german readers. PsyArXiv.

Sascha Schroeder, Kay-Michael Würzner, Julian Heister, Alexander Geyken, and Reinhold Kliegl. 2015. childlex: A lexical database of german read by children. *Behavior research methods*, 47:1085–1094.

Shuang Song, Mengmeng Su, Cuiping Kang, Hongyun Liu, Yuping Zhang, Catherine McBride-Chang, Twila Tardif, Hong Li, Weilan Liang, Zhixiang Zhang, et al. 2015. Tracing children's vocabulary development from preschool through the school-age years: An 8-year longitudinal study. *Developmental science*, 18(1):119–131.

Lukas Thoma, Ivonne Weyers, Erion Cano, Stefan Schweter, Jutta L Mueller, and Benjamin Roth. 2023. Cogmemlm: Human-like memory mechanisms improve performance and cognitive plausibility of llms. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning, pages 180–185, Singapore. Association for Computational Linguistics.

Michael Tomasello. 2005. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Fiona J Tweedie and R Harald Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32:323–352.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Ludo Verhoeven, Jan van Leeuwe, and Anne Vermeer. 2011. Vocabulary growth and reading development across the elementary school years. *Scientific studies of reading*, 15(1):8–25.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, et al. 2023. Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

A Appendix

Description of picture story “Jenga” by DeepSeek model: *Das Bild zeigt Lea, Lars und Dodo, die zusammen Jenga spielen. Lea, das Mädchen mit den Zöpfen, zieht vorsichtig einen Holzblock aus dem Turm. Lars, der Junge mit dem Basecap, beobachtet gespannt, während Dodo, der blaue Vogel mit der Brille, aufgeregt daneben steht. Der Jenga-Turm steht auf einem Tisch und sieht schon ziemlich wackelig aus! Im Hintergrund sieht man ein Regal mit Spielen und Büchern. Die Wände sind bunt und es hängen gemalte Bilder. Alle sehen sehr konzentriert und ein bisschen nervös aus, weil der Turm gleich umfallen könnte. Es sieht nach viel Spaß aus!* (The picture shows Lea, Lars, and Dodo playing Jenga together. Lea, the girl with the braids, is carefully pulling a wooden block from the tower. Lars, the boy with the baseball cap, is watching intently, while Dodo, the blue bird with glasses, is standing excitedly beside them. The Jenga tower is on a table and already looks quite wobbly! In the background, you can see a shelf with games and books. The walls are colorful and there are painted pictures hanging. Everyone looks very focused and a bit nervous because the tower might fall over any second. It looks like a lot of fun!).