

Stories for the Win: Creating and Evaluating a Synthetic Dutch Children’s Storytelling Dataset

Sabijn Perdijk, Gijs Wijnholds, Suzan Verberne and M.J. van Duijn
University Leiden

Abstract

Stories have been shown to be potentially valuable for language model pretraining because their semantics are largely self-grounded. Existing human storytelling datasets are relatively small, and existing synthetic datasets lack realism. Here we describe the guided synthesis of a story corpus that is both large and realistic. We evaluate a set of metrics for measuring the quality of individual stories, including coherence, surprise, complexity, as well for quality of the dataset as a whole. Using the metrics we extensively evaluate multiple prompting approaches and three small models (3-8B). Based on the results, we steer dataset generation to yield a high-quality ‘synthetic twin’ of a Dutch children’s storytelling collection, similar in terms of lexical and semantic diversity, but much larger, which we call *Stories for the Win* (StoriesFTW). We validate our metrics and evaluate samples from our corpus with crowd, expert, and LLM judgements.¹

1 Introduction

Narratives shape society since the earliest beginnings: they explain, educate, and provide purpose to mankind (Koning, 2015). They play a constitutive role in child development, boosting acquisition of communicative skills (Southwood and Russell, 2004), world knowledge (McKeough and Genereux, 2003), and Theory of Mind (Nicolopoulou, 2018). Training of Large Language Models (LLMs) also builds upon narratives: novels and short stories are classic elements of pretraining datasets (e.g. Gao et al., 2020). So far it is largely unclear what exactly the contribution of narrative language is in the LLM pipeline, yet there are empirical indications that training on relatively small amounts of children’s stories can be an effective and data-efficient way to extract meaningful word

vectors (Van Dijk et al., 2023) or create small language models that are fluent in various domains (Eldan and Li, 2023; Paraskeva et al., 2025). This is in line with theoretical and empirical links between narrative, language learning, and cognitive development more broadly (see §2.1).

To enable further investigation of the role children’s narratives play in pretraining LLMs, **we construct a large and realistic synthetic narrative dataset**. In light of Wilcox et al. (2025)’s plea to use more data that resembles children’s input, and considering the theoretical and empirical work cited above and in §2.1, we focus on realistic everyday narratives for and by children. Existing child-narrative datasets do not match our criteria. TinyStories, a synthetic dataset by Eldan and Li (2023), is large enough, but its design provides a limited view of child narratives. Children Stories Text Corpus (CSTC) (Bens, 2021), a subset of Project Gutenberg, offers a collection of world literature for children, but the idiosyncratic stories it contains (e.g. fairy tales) mostly do not align with children’s daily experiences. Examples of datasets that do align with children’s worlds are the English Nicolopoulou Corpus on CHILDES (Nicolopoulou, 2019) and the Dutch ChiSCor (Van Dijk et al., 2023). The Nicolopoulou Corpus contains stories told by and for children in their daily environments, yet from a limited range of age groups and social backgrounds. ChiSCor contains stories from diverse age groups and backgrounds, but with 619 stories and 74k tokens is too small to serve as a pretraining corpus. To combine the qualities of ChiSCor with pretraining scale, we generate a 1000 times larger synthetic twin of this Dutch children’s narrative corpus. We do so with a systematic approach to prompt design and model selection, guided by metrics of story quality at different levels. We discuss and test the alignment of the metrics with human judgements. These metrics, partly newly constructed and partly grounded in previ-

¹Dataset available at <https://huggingface.co/datasets/Sabijn/StoriesfortheWin>, code available at: <https://github.com/sabijn/stories4thewin>

ous research, measure coherence, creativity, and complexity at the story-level and overall quality and diversity at the dataset-level. Afterwards, the stories are evaluated with human judgements.

Our contributions are: (1) a set of metrics for measuring the quality of synthetic story data benchmarked on human judgements, which we use for (2) systematic evaluation of models and prompting approaches to steer the creation of (3) a synthetic Dutch children’s story dataset delivered for language model training and research.

2 Background

2.1 Storytelling and Child Development

Engaging with stories, either by producing or consuming them, urges the human brain to engage in a form of ‘cognitive pretend play’ and project an imagined, alternative reality, referred to as the *story-world* (Smith and Roopnarine, 2018; van Duijn, 2025). Objects, characters, and events gain meaning against the background of this imagined story-world, enabling a form of surrogate grounding (van Dijk, 2025). In other words, whereas most forms of discourse are grounded in a context outside the linguistic realm (e.g., a physical setting with its objects, participants, actions, etc.), stories are to a large degree grounded in themselves. Creating and understanding self-grounded discourse helps children develop their language skills (Nicolopoulou, 2019). This self-grounding likely also allows LLMs to learn new words more efficiently (Ma et al., 2023; van Dijk, 2025).

2.2 Evaluating at Narrative and Dataset Level

Training data quality is recognized as a crucial factor in the success of LLMs (Gunasekar et al., 2023; Wilcox et al., 2025). We focus on quality at the level of the individual dataset entries (the stories) and at a global level.

Story level Chhun et al. (2022) define six human criteria for ‘good’ narratives: *Relevance*, *Coherence*, *Surprise*, *Complexity*, *Empathy*, and *Engagement*. These criteria were used to rank several automatic metrics, and it was concluded that current metrics are insufficient to completely resemble human ratings. We, therefore, propose additional metrics based on their criteria, aiming to better reflect human experience.

Ismayilzada et al. (2024) evaluate creative short story generation in humans and LLMs based on diversity, novelty, surprise, and complexity. Tian

et al. (2024) evaluate narrative development and plot progression, concluding that LLMs fall short in discourse understanding.

Another field working with narrative quality is automatic storytelling. Prior research in this field mainly focussed on storytelling based on plot planning (Yao et al., 2019). Current work, related to the *Coherence* criterion, focusses on maintaining narrative order (Yao et al., 2019) or alterations on the narrative order in the form of flashbacks (Han et al., 2022; Lu et al., 2023).

Various evaluation criteria associated with high-quality narrative, such as linguistic complexity, are also applicable to generated text in general. Researchers have released open-source Python packages for measuring and extracting repetition in text (Shaib et al., 2025), benchmarking linguistic diversity in LLMs (Guo et al., 2025), or exploring precision and recall for the diversity of LLMs (Le Bronnec et al., 2024). Additional criteria include novelty (McCoy et al., 2023) and linguistic complexity (Muñoz-Ortiz et al., 2024).

Dataset level Various research lines have focused on measuring the global quality of artificially generated datasets (Li et al., 2023; Chen et al., 2023; Long et al., 2024). Dang and Verma (2024) have sought to create a taxonomy and standardize metrics, proposing five categories for measuring overall dataset quality: linguistic, semantic, anomalies, diversity, and classifier performance.

2.3 Dataset Generation

Dataset generation is an evolved form of data augmentation and can be applied at various text levels, from word-level to document-level (Zhou et al., 2024). At the **word level**, the most intuitive data augmentation method is to substitute words with their synonyms. A thesaurus like WordNet (Kolomiyets et al., 2011) or trained neural networks (Zhang et al., 2015) are the primary techniques for this, but off-the-shelf LLMs can also be used (Smilga and Alabiad, 2024). In addition, words can be substituted with words close in the embedding space (Marivate and Sefara, 2020) or sampled from a learned distribution (Kobayashi, 2018). In back-translation, generally operating at the **phrase level**, a piece of text is translated into a pivot language and back (Sennrich et al., 2016). **Document-level** methods can roughly be divided in work that trains or fine-tunes task-specific generative models (Gangal et al., 2022; Anaby-Tavor et al., 2020; Kumar

et al., 2020) and work that uses off-the-shelf models (Yoo et al., 2021; Sahu et al., 2022; Piedboeuf and Langlais, 2023). Methods in the second line of work are often prompting-based. Within the narrative domain, examples include the mentioned TinyStories, a dataset created via prompting containing stories for 3- and 4-year-olds (Eldan and Li, 2023), and SimpleStories, a similar synthetic dataset to TinyStories (Finke et al., 2025).

3 Methods: Metrics

The metrics for story quality are described in §3.1 and benchmarked on human judgments in §5.1. The metrics for dataset quality are described in §3.2.

3.1 Quantifying Story Quality

We use the six human criteria defined by Chhun et al. (2022) as a guideline for categorizing our metrics. Before calculating the metrics, text pre-processing is applied (Appendix A).

Basic Requirements encompass the criteria *Relevance*, *Coherence*, and *Grammaticality*. We argue that *Relevance* is a criterion that is automatically met, as we assume that instruction-tuned LLMs would be well enough able to follow the instructions in a generation prompt to stay within the relevant text domain. A manual check was performed confirming this argumentation. Following Purdy et al. (2018), we measure *Coherence* with local contextuality, defined as the average of the cosine similarities of every adjacent pair of sentences in the story. As an embedding model for the individual sentences, we use the SentenceTransformer variant of bert-base-dutch-cased.

We validate *Grammaticality*, following Purdy et al. (2018), as the log probability of POS-tag sequences in a reference corpus, the Corpus Gesproken Nederlands (van Eerten, 2007).

Surprise is defined as how surprising the end of the story was (Chhun et al., 2022) or the broader degree of deviation from an expected pattern (Karampiperis et al., 2014; Ismayilzada et al., 2024). To measure the degree of deviation from expected patterns, we propose a slightly adjusted form of perplexity: **Creative Perplexity**. This metric employs perplexity in combination with linguistically constrained bigrams to capture higher level creative constructs. An additional explanation of this metric is added in Appendix B.1.

Another way to measure surprise, is **diversity**. We measure diversity at the lexical and semantic levels. On a lexical level, we use Moving-Textual-Lexical-Diversity (MTLD) (McCarthy, 2005), the average segment length for segments with a given type-token ratio. On the semantic level, we employ Self-BLEU (Zhu et al., 2018).

Complexity describes how elaborate the story is. This can be measured at word level, i.e., lexical complexity, and sentence level, i.e., syntactic complexity. We approximate lexical complexity with two simple metrics well-established in readability research: the number of unique words in a story (UW) (Holmes, 1985) and the average word length (WL) (Flesch, 1948). The syntactic complexity is measured using the average occurrence of clausal complements (AC) (Kyle and Crossley, 2018), the mean dependency distance (DD) (Liu, 2008; Liu et al., 2017), and the average syntactic tree depth (SD) (Yngve, 1960).

Human evaluation is used to benchmark the automated metrics and measure more subjective story qualities such as *Engagement*. Sixty participants (recruited via Prolific) each rated ten stories, and three experts each rated thirty stories, sampled from a mix of StoriesFTW and ChiSCor, on a five-point Likert scale with the following items: *grammaticality*, *coherence*, *originality*, *creativity*, *complexity*, *likeability*, and *human likeness*. In total, two-hundred stories received 3 independent crowd ratings and thirty stories 3 expert ratings. For a detailed description of the human evaluation, see Appendix E.1.

LLM-as-a-judge is performed to investigate whether LLMs judge real and generated stories similarly to humans. Bound by computational resources and the ethical guidelines of ChiSCor, stating that the stories cannot be fed to a proprietary model, we chose the following four models: qwen2.5:14b (Yang et al., 2024), mistral-small13.2 (MistralAI, 2025), aya-expanse:32b (Dang et al., 2024), and gemma3:12b (Gemma Team et al., 2025). The setup is kept as similar to the human evaluation as possible. For a detailed description of the LLM-as-a-judge evaluation, see Appendix F.

3.2 Quantifying Dataset Quality

To measure the quality of the overall dataset, we focus on dataset diversity from the Dang and Verma

(2024) taxonomy (given that linguistic and semantic quality are already measured within the stories themselves). The four diversity metrics used are explained below:

MAUVE is a measure designed to compare generated and human text distributions, employing the Kullback-Leibler divergence in a quantized, low-dimensional space (Pillutla et al., 2021). A high MAUVE score indicates similar diversity and quality in the generated and human texts; a low score could indicate a lack of variety or poor quality in the generated text.

Vendi Score is a content-based diversity metric capturing semantic variations using sentence embeddings (Friedman and Dieng, 2023). It is defined as the exponential of the Shannon entropy of the eigenvalues of a similarity matrix and can be interpreted as the effective number of dissimilar elements in a story.

Compression Ratio is the ratio between the size of a compressed text and that of the original text (Shaib et al., 2025). A higher compression ratio implies a high number of redundant parts in the text. This metric, however, heavily relying on the size of the story, is sensitive to text length, complicating the interpretation of the final score.

Model-as-a-metric for Lexical Semantics To evaluate the lexical semantic properties of the generated data, we train a Word2Vec model and evaluate how the semantic space correlates with human judgments, by computing Spearman correlation against the Dutch Simlex999 dataset (Brans and Bloem, 2024). A high correlation indicates that the semantic space built by the candidate dataset is similar to the human semantic space, suggesting that this dataset contains sufficient information to train a model aligned with human preferences.

4 Methods: Dataset Generation

The dataset generation is divided into two phases: first, designing (§4.2) and evaluating (Appendix C.4) potential generation prompts and, after that, the generation of the dataset (§4.3). The model used for generation is described in §4.1.

4.1 Model Choice

We prompted gemma-3-4b (Gemma Team et al., 2025), phi-3.5-mini (Abdin et al., 2024), and llama-3.1-8b (Grattafiori et al., 2024) to each

generate 50 stories, which we analysed qualitatively and quantitatively. The set generated by llama-3.1-8b most closely resembled the language of children, so henceforward llama-3.1-8b is used. The three story sets and the analysis are added in Appendix C.2.

4.2 Designing Story Generation Prompts

The potential prompts are structured using the framework of Long et al. (2024), who divide a prompt into a *task*, a set of *generation conditions*, and *in-content demonstrations*.

Task Specification The goal of the research, generating stories, forms the basis of the task (hereafter, *baseline*). Additionally, previous research suggests that adding an in-domain persona to the task improves the overall quality of the generated answer (Salewski et al., 2023). So, in the light of our objective to match the narration style of ChiSCor, we investigate the impact of adding the persona of a child between 4-12 years old to the task (hereafter, *persona*).

Demonstrations We investigate whether few-shot prompting with three randomly sampled ChiSCor stories, improves the quality and diversity of the generated stories (referred to as *few-shot*).

Conditions Four different conditions to steer the generation space are considered. Previous research experimented with adding conditions for a theme, topic, style, narrative feature, grammar feature, and an author persona to steer the generation space (Finke et al., 2025). To avoid overly stylised stories, uncommon in the genre of children’s stories, we only investigate *themes* and *narrative features* by adjusting the list used by Finke et al. (2025) to contain only elements common to child language. The complete lists are shown in Appendix C.3.

Additionally, following Eldan and Li (2023), we evaluate the effect of adding three randomly sampled words: a noun, a verb, and an adjective. We build on their idea of sampling from a pre-defined dictionary by randomly selecting three words from ChiSCor stories (hereafter, *ChiSCor words*).

Finally, just as Finke et al. (2025), we observe that the majority of generated stories start with the genre-specific phrase ‘Once upon a time...’ (Dutch: ‘Er was eens...’). To minimise that amount, we investigate the effect of adding a compulsory POS tag at the beginning of the story. These POS tags are not randomly sampled but drawn from the POS-

	ChiSCor			StoriesFTW		
	Public	Expert	LLM	Public	Expert	LLM
Grammar	.22 [.07, .36]	.32 [-.05, .55]	.32 [.21, .41]	.12 [-.01, .24]	.56 [.25, .76]	.13 [.05, .22]
Coherence	.30 [.16, .42]	.49 [.05, .74]	.39 [.29, .48]	.19 [.05, .31]	.41 [-.01, .66]	.28 [.17, .38]
Originality	.32 [.20, .44]	.13 [-.18, .38]	.28 [.17, .37]	.05 [-.08, .18]	-.15 [-.37, .11]	-.10 [-.16, -.03]
Creativity	.41 [.29, .51]	.28 [-.04, .53]	.63 [.54, .70]	.05 [-.07, .16]	-.14 [-.39, -.04]	.24 [.13, .33]
Complex	.35 [.23, .46]	.41 [.11, .57]	.37 [.26, .46]	.09 [-.03, .20]	-.28 [-.39, .14]	-.01 [-.11, .07]
Likeability	.17 [.04, .31]	.48 [.13, .66]	.53 [.43, .61]	.10 [-.03, .22]	.12 [-.19, .36]	.28 [.14, .39]
Human	.05 [-.06, .18]	-.21 [-.36, -.04]	.04 [-.04, .11]	.02 [-.10, .13]	-.36 [-.42, -.31]	.09 [.09, .30]

Table 1: Krippendorff’s α per criterion and evaluation method for stories sampled from ChiSCor (left) and StoriesFTW (right). Brackets show 95% bootstrapped confidence intervals.

	Creativity	Gramm.	Coherence			Complexity			
	<i>C. PPL</i>	<i>Gramm.</i>	<i>Self-BLEU</i>	<i>LC</i>	<i>UW</i>	<i>WL</i>	<i>AC</i>	<i>DD</i>	<i>SD</i>
ρ_{public}	.181	.181	-.301*	-.154	.721*	.229	.122	.356*	.217
ρ_{expert}	.463	.339	-.523	.015	.846*	.403	.133	.033	-.009
ρ_{LLM}	.257*	.331*	-.154	-.089	.826*	.098	.304*	.327*	.230*

Table 2: Spearman’s ρ between the automated metrics and the corresponding human category. * means $p < 0.05$. *C. PPL*: Creative Perplexity; *LC*: Local Contextuality; *UW*: Unique Words; *WL*, average Word Length; *AC*: Average clausal Components; *DD*: Dependency Distance; *SD*: Syntactic tree Depth.

tag distribution of ChiSCor to mimic ‘real’ story openings (referred to as *POS-tags*).

4.3 Generating Stories for the Win

Stories for the Win (henceforward called *StoriesFTW*) is generated with the set of prompts that generated stories closest to ChiSCor, according to our analysis. The generation of these candidate stories is described in Appendix C.4. To avoid a collapse in the diversity and quality of the data, the dataset quality metrics, MAUVE and Vendi score, are calculated every 5000 stories, using a subset of 50 randomly sampled stories. If either metric drops by more than 20%, the prompt is changed to the next-best-performing prompt. The generation is performed on an RTX 3090, with a temperature of 0.7 and sampling disabled (GPU < 500 hours). The generation settings are chosen based on preliminary experiments.

To evaluate the resulting quality of StoriesFTW and keep the comparison with ChiSCor fair, we sample five times 600 stories and report the mean across runs.

5 Results

We first establish the usability of the metrics by correlating them with human and LLM judgements (§5.1) and showing that they are able to differentiate human and generated stories (§5.2). Next, we discuss which prompting methods generate a dataset most similar to ChiSCor (§5.3). Finally, we

compare StoriesFTW to ChiSCor (§5.4).

5.1 Validating the Metrics

The metrics are validated on human, both crowd and expert, and LLM judgements on stories sampled from ChiSCor. Firstly, we report inter-rater reliability scores (Krippendorff’s α (Krippendorff, 2018)) in Table 1, showing poor agreement across the board; *Human likeness* even shows a structural disagreement for the expert ratings ($\alpha = -.21$). The low scores can be attributed to the subjectivity of the task, a non-standardized annotation scheme, or a misalignment between adult expectations and children’s actual language use.

The scores per story are averaged and correlated with the automated metrics using Spearman’s ρ (Table 2). Overall, weak non-significant correlations are found, except for the correlation between *unique words* and *Complexity* ($\rho = .721$), which is not found in prior work (Cachola et al., 2025). A handful of metrics has a weak significant correlation with one of the evaluation methods, such as for *Creativity* ($\rho_{\text{LLM}} = .257$). Overall, the low α makes it difficult to pinpoint whether the low correlations are caused by malfunctioning automated metrics or unreliable human judgements. Low correlations with metrics well-grounded in the literature, such as syntactic depth, seem to point to unreliable human judgements. While remaining cautious, we decided to use all metrics in the generation process.

	Suprise	Diversity		Complexity				
	<i>C. PPL.</i>	<i>Self-BLEU</i>	<i>MTLD</i>	<i>UW</i>	<i>WL</i>	<i>AC</i>	<i>DD</i>	<i>SD</i>
<i>ChiSCor</i>	273	.160	30.3	58.4	3.70	.277	2.66	2.37
<i>Baseline</i>	179	.201	36.1	75.3	4.01	.172	2.42	2.28
<i>Persona</i>	228	.158	51.0	86.2	4.11	.281	2.80	2.72
<i>Few-shot</i>	198	.115	56.5	63.7	3.75	.416	2.96	2.83
<i>Themes</i>	242	.188	54.0	92.4	4.08	.303	2.83	2.77
<i>Narrative feat.</i>	348	.165	52.0	85.6	4.05	.377	3.01	2.60
<i>ChiSCor Words</i>	315	.160	53.1	85.9	4.14	.421	2.94	2.79
<i>POS-tags</i>	290	.145	53.8	85.7	4.08	.337	2.84	2.72

Table 3: Story quality results per prompting element. Bold-faced values indicates the value closest to ChiSCor on that metric. *C. PPL*: Creative Perplexity; *MTLD*: Measure of Textual Lexical Diversity; *UW*: Unique Words; *WL*, average Word Length; *AC*: Average clausal Components; *DD*: Dependency Distance; *SD*: Syntactic tree Depth.

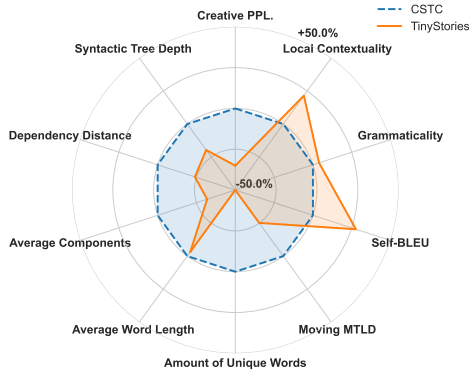


Figure 1: Relative scores of TinyStories compared to a natural story dataset (CSTC) on our defined metrics. Scores are capped on a 0.5 relative difference for visualisation purposes.

5.2 Differentiating Generated Stories

We use two existing datasets containing children stories, one human-written (a manual selection from CSTC; see Appendix A) and one synthetic (TinyStories), to show the metrics abilities. The comparison is visualised in Figure 1. Each axis represents one of the tested metrics. The dashed blue line represents scores for the CSTC corpus; the solid orange line represents the scores for TinyStories relative to the CSTC corpus. The figure shows that the synthetic corpus is similar to the human-written one in terms of *grammaticality* and *avg. word length*, but different in all other dimensions, showing that the metrics are able to measure a difference between human and generated stories.

5.3 Prompt Evaluation Results

To test which prompt conditions contribute to the diversity and quality of the generated dataset, we systematically evaluate all tasks, conditions, and demonstrations. Table 3 shows the influence of each addition to the baseline prompt on quality

	MAUVE	VS	CS
<i>ChiSCor</i>	1.00	24.4	3.15
<i>Baseline</i>	.749	7.35	3.30
<i>Persona</i>	.666	6.78	3.14
<i>Few-shot</i>	.914	12.7	2.78
<i>Themes</i>	.679	9.06	2.98
<i>Narrative feat.</i>	.674	10.4	2.93
<i>POS-tags</i>	.641	8.28	2.96
<i>ChiSCor words</i>	.641	11.4	3.15

Table 4: Dataset quality results per prompting element. Bold-faced values indicates the value closest to ChiSCor on that metric. MAUVE score is relative to ChiSCor. VS: Vendi Score; CS: Compression Score.

metrics at the story-level and Table 4 at the dataset-level.

5.3.1 Story Quality

Basic requirements The results for the coherence metric show little to no difference and a Kruskal-Wallis test, combined with Dunn’s test, only showed a statistically significant difference between ChiSCor and few-shot prompting ($H = 25.5$, $p < 0.05$). Despite the statistically significant difference, the effect sizes were both small, suggesting limited practical relevance. The same observation can be made for the results for the grammaticality metric; here the Kruskal-Wallis test showed no statistically significant difference between the datasets. The detailed results are in Appendix B.2. These results indicate that the coherence and grammaticality of the candidate datasets are roughly similar to the ones of human stories, supporting the decision to continue testing with all prompt additions.

Surprise, Diversity, Complexity Table 3 shows the results of the *Surprise*, *Diversity*, and *Complexity* metrics. First of all, it is visible that the POS-

	MAUVE	VS	CS	MaM
<i>ChiSCor</i>	1.00	24.4	3.15	.360
<i>StoriesFTW</i>	.612	17.3	3.01	.324

Table 5: Dataset quality results for ChiSCor and the new dataset. *MaM*: model-as-a-metric, measured as Spearman correlation with $p < 0.05$.

tags prompting element scores closest to ChiSCor for *creative perplexity*. Narrative features and ChiSCor words receive an even higher creativity score, indicating that these prompt elements give access to a larger creative generation space. More research is needed to confirm whether it is caused by genuine creativity or that is impacted by the lexical and syntactic complexity of the stories. Second, the baseline prompt achieves the most similar scores to ChiSCor on *lexical diversity* and *syntactic tree depth*. For the other complexity metrics persona prompting scores closest to ChiSCor. We hypothesize that the two task-only prompts, baseline and persona, score closest to ChiSCor because it does no additional request, allowing the model to focus on generating child language. Few-shot prompting scores close to ChiSCor on *lexical complexity*.

5.3.2 Dataset Quality

MAUVE, Vendi, and Compression ratio In Table 4, it is evident that few-shot prompting generates a dataset closest in diversity to ChiSCor. The high MAUVE score shows that it most closely resembles the ChiSCor distribution among the other prompting additions; a relatively unsurprising result given the ChiSCor demonstrations in the prompt. More surprisingly, baseline prompting also has a high MAUVE score, indicating that it is relatively similar to the ChiSCor distribution. The low compression score of few-shot prompting supports the claim that it creates a diverse dataset. Surprisingly, ChiSCor achieves one of the highest compression scores, suggesting it is not as diverse as previously assumed. This result can be explained by an inherent drawback of compression scores, which are lower when the texts are longer (Shaib et al., 2025). The length distribution of the generated texts are relatively equal, as they are generated with the same length condition of 200 words. ChiSCor, however, contains stories of varying lengths, making its compression score less robust.

Lexical Semantics Analysis We consider correlation scores with SimLex999 for the highest scor-

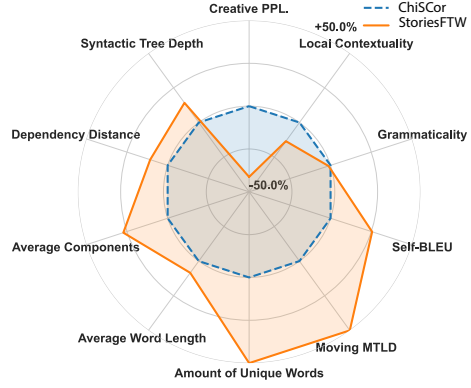


Figure 2: Relative scores of StoriesFTW compared to a natural story dataset (ChiSCor) on our defined metrics. Scores are capped on a 0.5 relative difference for visualisation purposes.

ing prompting elements on dataset quality. Surprisingly, the baseline prompt yields the same Spearman correlation as ChiSCor ($\rho = .360$), suggesting that the baseline stories are suitable for building a lexical-semantic space. The high score, however, is more likely due to the high frequency of common lemmas, which provide more context to build a meaningful word vector. Further, the absence of correlation for the ChiSCor words prompt ($\rho = .099$) is most likely due to the randomness introduced by the compulsory words, leading to words occurring in less likely contexts. Few-shot prompting shows a weak, non-significant correlation ($\rho = .200$). Table 7 in Appendix B.3 contains all correlation scores with their associated p-values.

In conclusion, the baseline prompting is closest to the syntactic complexity of ChiSCor and has a semantic space that better correlates with human judgements. Few-shot prompting generates the most diverse dataset and achieves a lexical complexity closest to ChiSCor. Further, ChiSCor words has a Self-BLEU score closest to ChiSCor and a relatively high *creative perplexity*. The prompting element closest to ChiSCor on *creative perplexity* is POS-tags. Therefore, we decide to start generating the with few-shot prompt, prioritising dataset diversity, and switch to ChiSCor words and POS-tags if the diversity during generation is dropping.

5.4 Analysing StoriesFTW

The final prompts and exemplary stories are added in Appendix D.

Automatic Metrics Figure 2 shows the relative differences between ChiSCor and StoriesFTW on the story quality metrics. The solid orange line

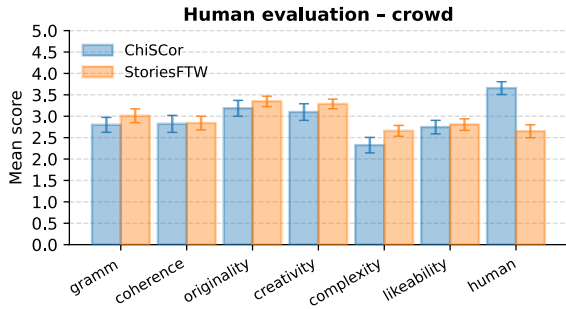


Figure 3: Mean scores calculated over stories on all criteria for crowd evaluation. Error bars indicate 95%-CI over stories. Gramm. is *Grammaticality*.

represents the relative scores of the new dataset compared to ChiSCor. The figure shows that the datasets correspond exactly on *grammaticality* and are close on *syntactic tree depth* and *dependency distance*. StoriesFTW has lower values on *creative perplexity* and *local contextuality*. The lower value for creative perplexity probably indicates fewer creative stories within StoriesFTW, but this conclusion needs to be drawn with caution as the creativity metric only weakly correlated with human judgements (§5.1). StoriesFTW scores higher on all lexical and syntactic diversity and complexity metrics, indicating that the language in the generated dataset is more complex and probably less child-like.

In Table 5, the results of the dataset metrics are shown for ChiSCor and StoriesFTW. The MAUVE score is relatively low compared to the score achieved during prompt testing. A low MAUVE score can be caused by low quality or low diversity. The relatively high Vendi Score and the lower compression score indicate that StoriesFTW should have a similar diversity to ChiSCor, suggesting that the lower MAUVE score is due to the lower quality of the generated dataset. Despite this, the SimLex score is comparable to ChiSCor’s, indicating that a meaningful semantic space can be constructed using StoriesFTW.

Human Evaluation The inter-rater reliability scores for StoriesFTW, reported in Table 1, point to poor agreement and in some cases even structural disagreement (Krippendorff, 2018), indicating that the task is highly subjective. Noticeable is that α is even lower than for the metric benchmarking (§5.1). Due to these low agreement scores, the results need to be interpreted with care, especially as averages of these ratings are not representing the variety of human judgements.

In Figure 3, the average crowd scores per criterion are plotted. Note that error bars indicate 95% confidence intervals across stories. The expert and LLM scores are added in Appendix E.2 and F.2. The Mann-Whitney U test show no statistical difference between StoriesFTW and ChiSCor on all criteria except for *Complexity* ($U = 3715$, $n_1 = n_2 = 100$, $p < 0.05$) and *Human likeness* ($U = 8205.5$, $n_1 = n_2 = 100$, $p < 0.05$). The statistical difference on complexity aligns with the difference found with the automated metrics.

6 General Discussion

While testing the metrics, the low observed correlation between the automated metrics and human evaluation was concerning. This can be caused by the subjective nature of the story rating task, already observed in prior work (Peeperkorn et al., 2025). This hypothesis is strengthened by the low inter-rater reliability for both the crowd and the expert ratings. This could be improved by adding more annotators (Long and Wang, 2022).

The low correlations can, however, not completely be attributed to the human evaluations. It is possible that the chosen metrics are not suited to measure children’s language as they are designed and benchmarked for adults. This could be the case for the creativity metric which is bound to grammatical models trained on adult text. Further, this explanation could shed light on the low correlation with the SimLex999 score for two of the four generated datasets as SimLex contains adult synonym judgements, and perhaps the lemma pairs should, in the opinion of children, receive a different score. Last, the low coherence scores for ChiSCor is also hypothesised to be inherent to the type of stories it contains: short spontaneous child narratives where events progress quickly. This results in relatively low similarity between adjacent sentences.

It is difficult to say whether our generated dataset StoriesFTW matches the creativity present in ChiSCor. This uncertainty aligns with prior work: some work concludes that generated stories are less creative than human stories (Ismayilzada et al., 2024), while others find no significant difference between generated and human children stories (Orwig et al., 2024). Besides that, we find, in line with prior work (Ismayilzada et al., 2024), that the generated stories are linguistically more complex than human stories. This difference, however, is not visible in TinyStories and CTSC. This is probably due to the

different ages of the target groups of TinyStories and CTSC.

The relatively low MAUVE score for our generated dataset deemed further exploration. A qualitative post-hoc inspection of the data indeed shows that there are several grammatical mistakes in the stories, such as the use of incorrect definite articles. This is most likely due to llama3.1-8b being multilingual, but not overly trained on Dutch. If these errors, however, are similar to those children make while learning the language, they would not be problematic in light of our aims. Errors young children make are usually associated with over-regularization, omissions, and word order errors (Brown, 1973). In the case of Dutch definite articles, monolingual Dutch children overgeneralize the common gender article ‘de’ to neuter nouns (Blom et al., 2008). This hints to the fact that at least some of the errors are driven by few-shot prompting with colloquial speech examples.

7 Conclusion

In this paper we have proposed and validated a set of metrics for story and dataset quality and we used the metrics for generating a large collection of stories (StoriesFTW) resembling a Dutch children’s storytelling collection (ChiSCor). We found that few-shot prompting with three demonstrations sampled from ChiSCor resulted in narrative language that most closely resembled ChiSCor. On the story level, we combined three prompting techniques to generate a diverse narrative dataset 1000 times ChiSCor’s size, amounting to 619,000 stories.

Our evaluation shows low correlation between the automated metrics and the human evaluation, which can be partly explained by the subjective nature of the task. Despite that drawback, we are carefully positive about our result, as the generated stories cannot be distinguished from ChiSCor stories on all defined categories except for *Complexity* and *Human likeness*.

For the near future, we hope that StoriesFTW enables further research into the role that narrative language can play in the LLM training pipeline. Second, additional approaches to enhancing creativity and diversity of generated narratives can be explored, such as setting up a co-creative writing process with children and LMs.

Ethical Considerations

For the human evaluation, we followed the checklist of the Ethics committee of our university, which indicated that evaluation by the ethics review committee was not needed, because we do not collect personal data, and do not involve participants from vulnerable user groups.

The use of LLMs has a considerable environmental footprint. However, as this research aims to present a dataset to improve pretraining efficiency, we deem this justified.

Limitations

The story quality analysis is restricted to metrics capturing lexical and syntactic complexity, while discourse-level complexity is not assessed. As a result, we cannot draw conclusions on the quality of higher-order structural properties of the generated stories, such as discourse organization. A second limitation concerning the metrics is that the notion of diversity is limited to *linguistic* diversity. Broader aspects of diversity, including societal representation and potential biases in generated stories, are not explicitly addressed. Thirdly, multiple metrics rely on SpaCy for syntactic and dependency parsing. During this research, we established that SpaCy’s parsing performance for Dutch is not impeccable, which may have led to parsing errors and, consequently, distorted values for metrics that depend on its syntactic analyses. Finally, the use of a relatively small language model may have introduced grammatical errors in the generated Dutch texts. Employing a larger language model could potentially reduce such errors and yield more linguistically robust outputs. It could, however, also lead to less childlike language.

References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *Preprint*, arXiv:2404.14219.
- Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Pro-*

- ceedings of the AAAI conference on artificial intelligence*, volume 34(05), pages 7383–7390.
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, and 1 others. 2025. Lms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255.
- Eden Bens. 2021. [Children stories text corpus](#).
- Elma Blom, Daniela Poliřenská, and Fred Weerman. 2008. Articles, adjectives and age of onset: The acquisition of dutch grammatical gender. *Second language research*, 24(3):297–331.
- Lizzy Brans and Jelke Bloem. 2024. [SimLex-999 for Dutch](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14832–14845, Torino, Italia. ELRA and ICCL.
- Roger Brown. 1973. *A first language: The early stages*. Harvard University Press.
- Isabel Cachola, Daniel Khashabi, and Mark Dredze. 2025. Evaluating the evaluators: Are readability metrics good measures of readability? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24022–24038.
- Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. [PLACES: Prompting language models for social conversation synthesis](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.
- Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. [Of human criteria and automatic metrics: A benchmark of the evaluation of story generation](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. [Aya expand: Combining research breakthroughs for a new multilingual frontier](#). *Preprint*, arXiv:2412.04261.
- Vu Minh Hoang Dang and Rakesh M Verma. 2024. Data quality in nlp: Metrics and a comprehensive taxonomy. In *International Symposium on Intelligent Data Analysis*, pages 217–229. Springer.
- Ronen Eldan and Yuanzhi Li. 2023. Tinstories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.
- Lennart Finke, Chandan Sreedhara, Thomas Doods, Mat Allen, Juan Rodriguez, Noa Nabeshima, Thomas Marshall, and Dan Braun. 2025. [Parameterized synthetic text generation with simplestories](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.
- Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.
- Dan Friedman and Adji Bousso Dieng. 2023. [The vendi score: A diversity evaluation metric for machine learning](#). *Transactions on Machine Learning Research*.
- Varun Gangal, Steven Y Feng, Malihe Alikhani, Teruko Mitamura, and Eduard Hovy. 2022. Nareor: The narrative reordering problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36(10), pages 10645–10653.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. [The pile: An 800gb dataset of diverse text for language modeling](#). *Preprint*, arXiv:2101.00027.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, and 1 others. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.
- Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526.
- Rujun Han, Hong Chen, Yufei Tian, and Nanyun Peng. 2022. Go back in time: Generating flashbacks in stories with event temporal prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1450–1470.

- David I Holmes. 1985. The analysis of literary style—a review. *Journal of the Royal Statistical Society: Series A (General)*, 148(4):328–341.
- Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2024. Evaluating creative short story generation in humans and large language models. *arXiv preprint arXiv:2411.02316*.
- Pythagoras Karampiperis, Antonis Koukourikos, and Evangelia Koliopoulou. 2014. Towards machines for measuring creativity: The use of computational tools in storytelling activities. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, pages 508–512. IEEE.
- Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457.
- Oleksandr Kolomiyets, Steven Bethard, and Marie Francine Moens. 2011. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 271–276.
- Hugo H. Koning. 2015. *Mythologie. Elementaire deeltjes*; 32. AUP, Amsterdam.
- Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd workshop on life-long learning for spoken language systems*, pages 18–26.
- Kristopher Kyle and Scott A Crossley. 2018. Measuring syntactic complexity in l2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2):333–349.
- Florian Le Bronnec, Alexandre Verine, Benjamin Nègrevergne, Yann Chevalere, and Alexandre Alauzen. 2024. [Exploring precision and recall to assess the quality and diversity of LLMs](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11418–11441, Bangkok, Thailand. Association for Computational Linguistics.
- Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. [Synthetic data generation with large language models for text classification: Potential and limitations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.
- Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.
- Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of life reviews*, 21:171–193.
- Haiying Long and Jue Wang. 2022. Dissecting reliability and validity evidence of subjective creativity assessment: A literature review. *Educational psychology review*, 34(3):1399–1443.
- Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082.
- Zhicong Lu, Li Jin, Guangluan Xu, Linmei Hu, Nayu Liu, Xiaoyu Li, Xian Sun, Zequn Zhang, and Kaiwen Wei. 2023. Narrative order aware story generation via bidirectional pretraining model with optimal transport reward. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6274–6287.
- Ziqiao Ma, Jiayi Pan, and Joyce Chai. 2023. [World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 524–544, Toronto, Canada. Association for Computational Linguistics.
- Vukosi Marivate and Tshephisho Sefara. 2020. Improving short text classification through global augmentation methods. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 385–399. Springer.
- Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.
- R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. [How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN](#). *Transactions of the Association for Computational Linguistics*, 11:652–670.
- Anne McKeough and Randy Genereux. 2003. [Transformation in narrative thought during adolescence: The structure and content of story compositions](#). *Journal of Educational Psychology*, 95(3):537.
- MistralAI. 2025. Mistral-small-3.2-24b-instruct-2506. <https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506>.
- Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265.

- Ageliki Nicolopoulou. 2018. Pretend and social pretend play: Complexities, continuities, and controversies of a research field. In *The Cambridge Handbook of Play*, pages 183–199. Cambridge University Press.
- Ageliki Nicolopoulou. 2019. Chapter 12. using a storytelling/story-acting practice to promote narrative and other decontextualized language skills in disadvantaged children. In *Narrative, Literacy and Other Skills*, volume 25, pages 263–284. John Benjamins Publishing Company.
- William Orwig, Emma R Edenbaum, Joshua D Greene, and Daniel L Schacter. 2024. The language of creativity: Evidence from humans and large language models. *The Journal of creative behavior*, 58(1):128–136.
- Andreas Paraskeva, Max Johannes van Duijn, Maarten de Rijke, Suzan Verberne, and Jan N van Rijn. 2025. [Data efficient pre-training for language models: An empirical study of compute efficiency and linguistic competence](#). In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*.
- Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2025. Mind the gap: Conformative decoding to improve output diversity of instruction-tuned large language models. *arXiv preprint arXiv:2507.20956*.
- Frédéric Piedboeuf and Philippe Langlais. 2023. Is chatgpt the ultimate data augmentation algorithm? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15606–15615.
- Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.
- Christopher Purdy, Xinyu Wang, Larry He, and Mark Riedl. 2018. Predicting generated story quality with quantitative measures. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 14(1), pages 95–101.
- Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57.
- Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models’ strengths and biases. *Advances in neural information processing systems*, 36:72044–72057.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 86–96.
- Chantal Shaib, Venkata S Govindarajan, Joe Barrow, Jiuding Sun, Alexa Siu, Byron C Wallace, and Ani Nenkova. 2025. Standardizing the measurement of text diversity: A tool and comparative analysis. In *Proceedings of The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 36–46.
- Veronika Smilga and Hazem Alabiad. 2024. Tüduo at semeval-2024 task 2: Flan-t5 and data augmentation for biomedical nli. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 737–744.
- Peter K Smith and Jaipaul L Roopnarine. 2018. *The Cambridge handbook of play: Developmental and disciplinary perspectives*. Cambridge University Press.
- Frenette Southwood and Ann F. Russell. 2004. [Comparison of Conversation, Freeplay, and Story Generation as Methods of Language Sample Elicitation](#). *Journal of Speech, Language and Hearing Research*, 47(2):366–376.
- Agnes Tellings, Nelleke Oostdijk, Iris Monster, Franc Grootjen, and Antal Van den Bosch. 2018. Basiscript: A corpus of contemporary dutch texts written by primary school children. *International journal of corpus linguistics*, 23(4):494–508.
- Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. [Are large language models capable of generating human-level narratives?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.
- Leonoor van der Beek, Gosse Bouma, Rob Malouf, and Gertjan Van Noord. 2002. The alpino dependency treebank. In *Computational linguistics in the Netherlands 2001*, pages 8–22. Brill.
- B.M.A. van Dijk. 2025. Theory of mind in language, minds, and machines: a multidisciplinary approach.
- Bram Van Dijk, Max van Duijn, Suzan Verberne, and Marco Spruit. 2023. Chiscor: A corpus of freely-told fantasy stories by dutch children for computational linguistics and cognitive science. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 352–363.
- Max J van Duijn. 2025. Introduction to part 3: Narrative meaning-making. *Social Psychology and the Ancient World: Methods and Applications*, 8:175.
- Laura van Eerten. 2007. Over het corpus gesproken nederlands. *Nederlandse Taalkunde*, 12(3):194–215.

Ethan Gotlieb Wilcox, Michael Y. Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of memory and language*, 144:104650–.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. [Qwen2 technical report](#). *Preprint*, arXiv:2407.10671.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33(01), pages 7378–7385.

Victor H Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Yue Zhou, Chenlu Guo, Xu Wang, Yi Chang, and Yuan Wu. 2024. A survey on data augmentation in large model era. *arXiv preprint arXiv:2401.15422*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Taxygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A Preprocessing

Each story is preprocessed using SpaCy’s pipeline (Dutch model: `n1_core_news_lg`; English model: `en_core_web_sm`): all words are lowercased, punctuation is removed, and the story is split into sentences. As the Dutch SpaCy does not handle multiple newlines or newlines combined with white spaces well, all newlines are replaced with a single newline.

Further, the CSTC contained a few works not conventionally seen as child literature. We manually removed these stories. The list we retained can be found in the code.

B Metrics

B.1 Additional Explanation Creative PPL.

Perplexity is the de facto evaluation metric of LMs; the lower the perplexity score, the better the model learned the data. Inverting this reasoning, a high perplexity would indicate that the model did not accurately learn the data distribution, implying that this type of story was not overly present in the training set and can therefore be considered creative or surprising.

Instead of using regular bigrams to represent the stories, we use linguistically constrained bigrams (i.e., bigrams selected based on topic-heavy dependency relations; for example, subject-verb relations), rather than regular bigrams, to create a representation at the discourse level rather than the word level. For example, in the sentence ‘The banana flew to Canada’, the selected bigram is ‘(subject, verb), (banana, fly)’. To capture the essence of the expected storyline, we create a linguistically constrained bigram lexicon of BasiScript ([Tellings et al., 2018](#)), a dataset containing text written by children in an open domain.

B.2 Results per Metric for Basic Requirements

In Table 6, the basic requirements metrics for story quality per prompting elements are shown.

B.3 Spearman correlation SimLex999

The spearman correlation with SimLex999 judgments for candidate datasets is shown in Figure 7.

	Coherence	Gramm.
<i>ChiSCor</i>	0.357	-2.15
<i>Baseline</i>	0.358	-2.19
<i>Persona</i>	0.333	-2.18
<i>Few-shot</i>	0.309	-2.11
<i>Themes</i>	0.338	-2.16
<i>Narrative feat.</i>	0.348	-2.15
<i>POS-tags</i>	0.336	-2.17
<i>ChiSCor words</i>	0.352	-2.15

Table 6: Basic requirement metrics for story quality per prompting element

	Spearman ρ	p -value
<i>ChiSCor</i>	.360	.02
<i>Baseline</i>	.360	.04
<i>Few-shot</i>	.200	.19
<i>ChiSCor words</i>	.099	.52

Table 7: Spearman correlation between fruitful prompting elements and SimLex999-score (Brans and Bloem, 2024). As the set of lemmas present in all datasets is too small to analyse (23), we sampled 45 lemmas for each dataset.

C Generation

C.1 Used prompts

The baseline prompt in English is shown in Figure 4 (translation in Figure 6). The adjusted prompt in English is shown in Figure 5 (translation in Figure 7).

System prompt You are the narrator of a short story (around 200 words).
Use only very simple words that a 3-year-old can understand.
Do not give the story a title or introduction.
User prompt Tell a story.

Figure 4: Baseline prompt in English

C.2 Model Choice

In Table 8, the scores on lexical and syntactical complexity for the three different models are shown. We chose to focus on the complexity metrics for model choice as it is most important that the model outputs language resembling ChiSCors language form.

It is immediately noticeable that all values of the

System prompt You are the narrator of a short story (around 200 words).
You are a child between 4 and 6 years old, and you are telling a story to classmates.
Your audience consists of children your age.
Do not give the story a title or introduction.
User prompt Tell a story.
> INSERT CONDITION OR DEMONSTRATION
SPECIFIC USER PROMPT.

Figure 5: Adjusted prompt in English.

System prompt Je bent een verteller van een kort verhaal (rond de 200 woorden).
Gebruik alleen hele simpele woorden die een 3-jarig kind kan begrijpen.
Geef het verhaal geen titel of introductie.
User prompt Vertel een verhaal.

Figure 6: Baseline prompt in Dutch.

Phi3.5 stories are too high to resemble the stories in ChiSCor and child language in general. Both the stories generated by gemma3 and llama3.1-8b receive scores close to ChiSCor, making a closer, qualitative look required. Below are the beginnings of two stories generated respectively by gemma3 and llama3.1-8b.

“Hey, listen up everyone! It was really great! I was in the woods, right? With my stuffed animal, Pip. Pip is a bunny, and he’s really smart. We went looking for a fairytale creature. I saw a squirrel first!...” - Gemma3

	Complexity				
	<i>UW</i>	<i>WL</i>	<i>AC</i>	<i>DD</i>	<i>SD</i>
<i>ChiSCor</i>	58.4	3.70	.277	2.66	2.37
<i>Llama3.1</i>	87.0	3.96	.238	2.83	2.61
<i>Phi3.5</i>	146	4.79	.360	3.52	3.86
<i>Gemma3</i>	111	3.89	.104	2.73	2.01

Table 8: Lexical and syntactic complexity results per tested model. Bold-faced values are the prompting elements that scored closest to ChiSCor on that metric. Further, the column names are as follows: *UW*, Unique Words; *WL*, average Word Length; *AC*, Average clausal Components; *DD*, Dependency Distance; *SD* Syntactic tree Depth.

System prompt Je bent een verteller van een kort verhaal (rond de 200 woorden).
 Je bent een kind tussen de 4 en 6 en je vertelt een verhaal aan klasgenoten.
 Je publiek bestaat uit kinderen van jouw leeftijd.
 Geef het verhaal geen titel of introductie.
User prompt Vertel een verhaal.
 > INSERT CONDITION OR DEMONSTRATION SPECIFIC USER PROMPT.

Figure 7: Adjusted prompt in Dutch.

“Hé, luister allemaal! Het was echt super! Ik was in het bos, hè? Met mijn knuffel, Pip. Pip is een konijntje en hij is heel slim. We gingen op zoek naar een sprookjesbeest. Ik zag eerst een eekhoorn!...”

“One time, I was at the playground with my friend Lotta. We were climbing on the big jungle gym when suddenly a rabbit popped out of its hole. It was as big as a dog!...” - Llama3.1-8b

“Een keer was ik met mijn vriendin Lotta in de speeltuin. We waren aan het klimmen in de grote klimrek en opeens kwam een konijn uit zijn holletje. Hij was zo groot als een hond! ...”

Both stories are written in language that resembles the simplicity of child language. Noticeable is that the three lines of the gemma story already contain three exclamation marks and a further investigation of the other stories generated by gemma show that all stories contain an abundance of exclamation marks. This way of storytelling feels, after reading multiple stories, quickly forced and generic, bringing us to the conclusion that llama3.1-8b generates the best stories for our purpose.

C.3 List of narrative elements and themes

Narrative elements: in medias res, een morele les, een onverwachte wending, een onbetrouwbare verteller, vooruitwijzing, innerlijke monoloog, symboliek, een niet-lineaire tijdlijn, een omgekeerde tijdlijn, circulaire verhaalsstructuur, een flashback, een geneste structuur, meerdere perspectieven, een cliffhanger, contrast (juxtapositie), climax-structuur

Story theme: sprekende dieren, fantasiewerelden, tijdreizen, een deadline of tijdslimiet, ruimteverkenning, mystieke wezens, onderwateravonturen, dinosaurussen, piraten, superhelden, sprookjes, het heelal, verborgen schatten, magische landen, betoverde bossen, geheime genootschappen, robots en technologie, sport, schoolleven, vakanties, culturele tradities, magische voorwerpen, verloren beschavingen, ondergrondse werelden, vervlogen tijdperken, onzichtbaarheid, reusachtige wezens, miniatuurwerelden, ontmoetingen met buitenaardse wezens, behekste plekken, vormverandering, eilandavonturen, ongewone voertuigen, geheime missies, droomwerelden, virtuele werelden, raadsels, rivaliteit tussen broers en zussen, schattenjachten, sneeuwavonturen, seizoenswisselingen, mysterieuze kaarten, koninkrijken, levende objecten, tuinen, verloren steden, de kunsten, de hemel

C.4 Evaluating Potential Prompts

For each prompt adjustment, 50 stories are generated with a temperature of 0.7 and sampling disabled. These stories are then evaluated with the metrics discussed in §3. The generation settings are chosen based on preliminary experiments.

All adjustments, except for the *baseline*, use a system prompt *including persona prompting* to keep the generation conditions as close as possible to the final prompt. The baseline and adjusted prompts in English are added together with the original Dutch ones in Appendix C.1.

The 200-word limit in both system prompts is based on the average length of stories in ChiSCor. In the baseline prompt, the target age corresponds to the one used in TinyStories to enable comparison. The target age in the adjusted system prompt is experimentally chosen (see Appendix C.5). The last line in both prompts is added post-hoc as models tended to generate a title or an introduction.

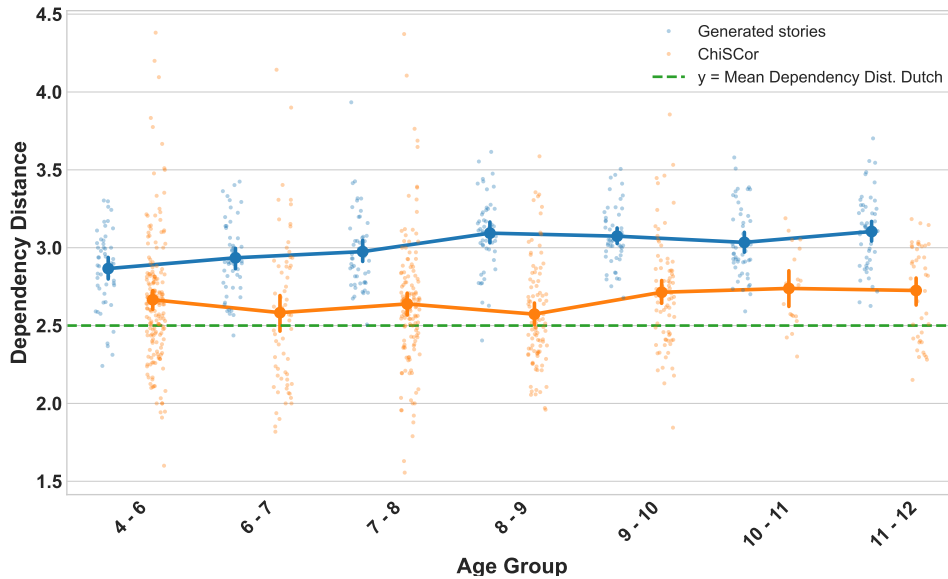


Figure 8: Dependency distance per age group. The blue and orange dots represent respectively the raw data of the generated stories and ChiSCor. The coloured bold lines represent the means of both datasets. Further, the dashed green line represents the mean dependency distance of the Dutch language, calculated by (Liu, 2008). Note that, this mean is calculated based on the Alpino dependency treebanks, containing mostly newspaper articles (van der Beek et al., 2002).

C.5 The Impact of Age in the Prompt

In the final prompt used to create the dataset with, the age group was kept constant between 4 and 6 even though the age group of ChiSCor ranges from 4 to 12. In Figure 8, we show the reasoning behind this choice. It shows that the syntactic complexity is almost stable across age groups as depicted with the solid blue line. Similarly, varying the age groups in the prompt shows a neglect-able increase in dependency distance throughout the age groups. In our effort to model the generated dataset like ChiSCor, we, therefore, chose to keep the age group constant while generation the new dataset.

System prompt You are the narrator of a short story (around 200 words).

You are a child between 4 and 6 years old, and you are telling a story to classmates.

Your audience consists of children your age.

Do not give the story a title or introduction.

User prompt Tell a story.

Here are 3 stories as inspiration:

Story 1: STORY_1

Story 2: STORY_2

Story 3: STORY_3

Figure 9: Final few-shot prompt.

D StoriesFTW

D.1 Used Prompts StoriesFTW

In Figure 9, the final few-shot prompt is shown (Dutch translation in Figure 10). In Figure 11, the final ChiSCor-words prompt is shown (Dutch translation in Figure 12). In Figure 13, the final POS-tag prompt is shown (Dutch translation in Figure 14).

D.2 Example Stories StoriesFTW

Story 1 “Once upon a time there was a girl and she had a friend. And she said: ‘Did you know what was really bad once?’ My father was at a festival once. And he had to go to the toilet. And then my father wanted to go to the toilet very quickly.

But there was no toilet nearby. So then he really needed to pee. But then he ended up peeing in his pants. So because of that, that girl told the friend... and then the other friend went and told her parents. To her teacher and to all the children in the whole city! And then the friend got very angry!”

Story 1 “Er was eens een meisje en die had een vriendin. En die zei: “Wist je wat heel erg was een keer?” Mijn vader was een keer op een festival. En hij moest toen naar de wc. En toen wou m’n vader heel snel naar de wc gaan. Maar er was geen wc in de buurt. Dus toen moest hij heel nodig plassen. Maar toen plaste die uiteindelijk in

System prompt Je bent een verteller van een kort verhaal (rond de 200 woorden). Gebruik alleen hele simpele woorden die een 3-jarig kind kan begrijpen. Geef het verhaal geen titel of introductie.

User prompt Vertel een verhaal. Hier zijn 3 verhalen om inspiratie uit op te doen:

Verhaal 1: STORY_1
Verhaal 2: STORY_2
Verhaal 3: STORY_3

Figure 10: Final few-shot prompt in Dutch.

User prompt Tell a story.
The story needs to contain the following verb: VERB, the following noun: NOUN and the following adjective: ADJECTIVE."

Figure 11: Final ChiSCor-words user prompt.

zijn broek. Dus daardoor zei dat meisje dat tegen de vriendin... en toen ging die andere vriendin het aan haar ouders vertellen. Aan haar juffrouw en aan alle kinderen in de hele stad! En toen werd de vriendin heel boos!"

Story 2 "First, that stupid chicken! It saw a snail and was so scared that it ran away... And then it laid an egg! How strange, isn't it? And then the snail ate it. That is very strange, too. I think I'd rather not talk about those animals. My girlfriend has a pet and it eats happily out of her hand. Now I have a story too. Once upon a time, there lived a boy in a world with knights. He wasn't very rich, but he did have many friends. In the forest, he saw a lot of animals: deer, squirrels, rabbits... And then he saw a broken house. But no one came out. Still, he went inside and there sat a grandma. She asked him if he wanted a candy. No, said the boy. But his own house caught fire later! What is this crazy story? He went back to the forest and asked the grandma: "What happened to my house?" And she answered: "You didn't eat my candy, now your house is in ashes."

Story 2 "Eerst die domme kip! Die zag een slak en schrok zo erg dat hij wegrende... En toen legde hij een ei! Wat vreemd, hè? En toen at de slak hem op. Dat is ook heel vreemd. Ik denk dat ik liever niet over die dieren praat. Mijn

User prompt Vertel een verhaal.
Het verhaal moet het volgende werkwoord bevatten: VERB, het volgende zelfstandig naamwoord: NOUN en het volgende bijvoegelijk naamwoord: ADJECTIVE."

Figure 12: Final ChiSCor-words user prompt in Dutch.

User prompt Tell a story.
Start the story with a word with the following pos-tag POS-TAG

Figure 13: Final POS-tag user prompt.

vriendin heeft een huisdier en die eet lekker uit haar hand. Nu heb ik ook een verhaal. Er woonde eens een jongen in een wereld met ridders. Hij was niet erg rijk, maar hij had wel veel vrienden. In het bos zag hij heel veel dieren: herten, eekhoorns, konijnen... En toen zag hij een kapot huis. Maar er kwam niemand uit. Toch ging hij naar binnen en daar zat een oma. Zij vroeg hem of hij een snoepje wilde. Nee, zei de jongen. Maar zijn eigen huis stond later in brand! Wat is dit gekke verhaal? Hij ging terug naar het bos en vroeg aan de oma: "Wat is er met mijn huis gebeurd?" En die antwoordde: "Je hebt mijn snoepje niet opgegeten, nu ligt je huis in as."

E Detailed Description of Human Evaluation

E.1 Set-up Human Evaluation

Questionnaire Set-up The frame of the questionnaire consisted of a consent question (Figure 15), information about the task and criteria and a repeating block of the story, a comprehension question, and a five-point Likert scale. The information about the task and criteria is shown in Figure 16. In the repeated block, first only the story was shown to force participants to read carefully. After that, a comprehension question was asked. The amount of wrong answers over all ten stories was counted as a proxy for individual judge reliability. When the comprehension question was correctly answered, the Likert items were shown.

Criteria The criteria *Grammaticality*, *Coherence*, and *Complexity* map directly to the criteria defined in §3. Then, the criterion *Likeability*

User prompt Vertel een verhaal.
Begin het verhaal met een woord met het
volgende pos-tag POS-TAG

Figure 14: Final POS-tag user prompt in Dutch.

By clicking "I consent" below, you confirm that you have read and agree to the following information.

- I confirm that I have been informed about the nature and method of the research.
- I agree to participate in this research. I understand that I can withdraw from my participation at any time.
- If my research results are used in scientific publications or made public in any other way, this will be done anonymously. If you experience accessibility issues during the experiment, please feel free to contact the researcher.

Figure 15: Consent question.

is added to measure *Engagement* (as defined by (Chhun et al., 2022)). We stray here from the original definition to make it more suitable for children’s stories. The criteria *Creativity* and *Originality* are respectively used to measure creativity and as construct validity. Last, the criterion *human likeness* is straight up used to investigate how close the generated stories align to the real stories.

In the questionnaire, the criteria are sparsely explained to participants to avoid a bias towards our own definition of the more subjective criteria, e.g. creativity. The explanation for the participants is shown in Figure 17.

Participants We recruited 60 participants with a Dutch nationality to rate in total 200 stories of which 100 from ChiScor and 100 from *StoriesFTW*. Each participant rated 10 stories such that all stories have 3 ratings. Participants were paid £12 per hour, which is the upper limit of the advised pay rate on Prolific. Besides that, we ask 3 experts (people with a PhD in literary studies) to rate 30 stories.

For this study, we ask you to evaluate 10 short stories based on 7 criteria: *Grammaticality*, *Coherence*, *Originality*, *Creativity*, *Complexity*, *Likeability*, and *Human likeness*.

The stories average around 200 words. First, only the story is shown for 15 seconds. This gives you time to read it carefully. Afterwards, a simple comprehension question is asked. After correctly answering the comprehension question, the criteria are displayed. You evaluate the criteria by giving a score on a scale of 1 to 5.

Some stories were told by children, others were generated by AI.

Figure 16: Information about the task and criteria.

E.2 Results of the Human Evaluation

The inter-rater reliability is divided based on source (either ChiScor or *StoriesFTW*) and are shown in Table 1.. The spearman correlations with the automatic metrics is shown in Table 2. The average score per criteria for the experts and LLMs are shown in Figure 18.

F Detailed Description of LLM-as-a-judge Evaluation

F.1 Set-up LLM Evaluation

The LLM-as-a-judge evaluation is modelled after the human evaluation. We analyse both the generation and the log probabilities. The prompts are given in Figure 19 and 20. We used the following four models: qwen2.5:14b (Yang et al., 2024), mistral-small3.2 (MistralAI, 2025), aya-expanse:32b (Dang et al., 2024), and gemma3:12b (Gemma Team et al., 2025) as judges.

F.2 Results LLM-as-a-judge Evaluation

The inter-rater reliability scores (Krippendorff’s α), reported in Table 1 point to poor agreement and in some cases even structural disagreement (Krippendorff, 2018), indicating that the task is highly subjective for LLMs.

In Figure 18, the average llm scores per criterion are plotted on the left side of the figure. The plotted values are the ratings based on the generations but it is confirmed that they align with the

To begin, we provide an explanation of the seven criteria:

1. Grammaticality. To what extent is the story grammatically correct? Do not be too strict: some sentence structures work in spoken language but not in written language.
2. Coherence. To what extent is the story a coherent whole? Think, for example, of a consistent storyline or the sudden introduction of characters.
3. Originality. How original do you find this story? There is no right or wrong answer here; we are looking for your opinion!
4. Creativity. Is this story creative in your eyes? There is no right or wrong answer here; we are looking for your opinion!
5. Complexity. How complex is this story? Think, for example, of the complexity of the sentence structure, but also that of the storyline.
6. Likeability. Did you enjoy reading this story? There is no right or wrong answer here; we are looking for your opinion!
7. Human likeness. How likely do you think that this story was generated by AI rather than told by a child?

2021), CGN (van Eerten, 2007), and BasiScript (Tellings et al., 2018). Tinystories is under a Community Data License Agreement – Sharing, Version 1.0 and CSTC under a CC0: Public Domain license. For the use of both CGN and BasiScript, a non-commercial license is signed. This license allows us to use the datasets for our own research and use it to develop new datasets as long as the datasets are not solely copied. With the creation of StoriesFTW, we adhere to this license.

Figure 17: Explanation for participants.

log-probabilities. Mann-Whitney U tests showed a statistical difference on all criteria ($532.5 < U < 6803$, $n_1 = n_2 = 100$, $p < 0.05$) except for *Originality* ($U = 5048.5$, $n_1 = n_2 = 100$, $p = 0.905$). The error bars are considerably smaller than for the expert ratings as the sample size is larger and the LLM evaluation has 4 raters per story. Interestingly, LLMs rated *StoriesFTW* higher on all criteria, except for *Human likeness*, confirming prior research (Bavaresco et al., 2025) that LLM might be biased towards generated text.

G License of Used Datasets

Within this work, we used the following datasets: TinyStories (Eldan and Li, 2023), CSTC (Bens,

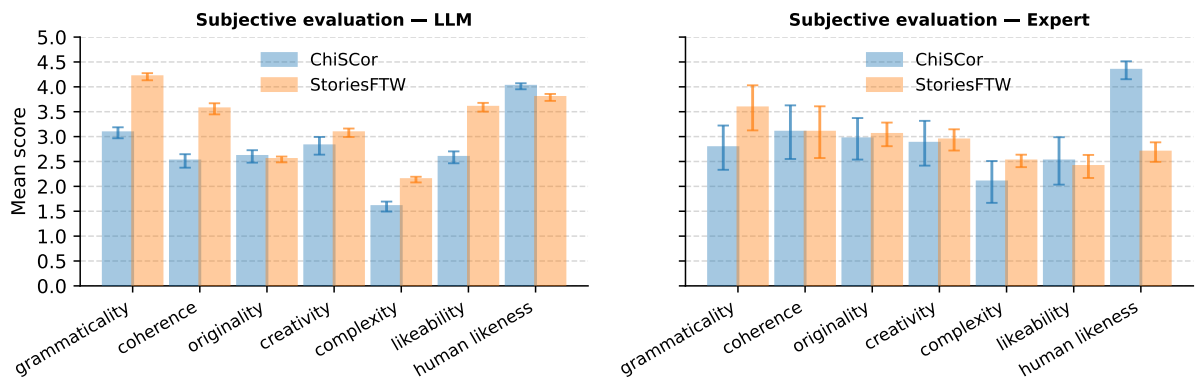


Figure 18: Mean scores on all criteria for public and expert human evaluation. Error bars indicate 95%-CI.

(Generation) System prompt You are an expert evaluator of Dutch children's stories. Your task is to rate a short Dutch story on seven quality criteria, each on a scale from 1 (very poor) to 5 (excellent).

The stories are a mix of real stories told by children and AI-generated stories designed to mimic child-told stories. Rate each story objectively based on the criteria below.

> INSERT CRITERIA AND EXPLANATIONS.

Return ONLY a JSON object with exactly these keys: grammaticality, coherence, originality, creativity, complexity, likeability, human likeness. Each value must be an integer between 1 and 5. Do not include any other text.

Example:

```
{"grammaticality": 4, "coherence": 3, "originality": 2, "creativity": 3, "complexity": 2, "likeability": 4, "human likeness": 3}
```

(Log probability) System prompt You are an expert evaluator of Dutch children's stories.

You will rate a story on a single criterion on a scale from 1 (very poor) to 5 (excellent).

Respond with ONLY a single digit: 1, 2, 3, 4, or 5. No explanation, no other text.

Figure 20: Prompt LLM evaluation based on log probabilities.

Figure 19: Prompt LLM evaluation.