MIRRORSTORIES: Reflecting Diversity through Personalized Narrative Generation with Large Language Models

Sarfaroz Yunusov, **Hamza Sidat**, and **Ali Emami**
Brock University, St. Catharines, Canada
{zw22fi, hs18so, aemami}@brocku.ca

Abstract

This study explores the effectiveness of Large Language Models (LLMs) in creating personalized "mirror stories" that reflect and resonate with individual readers' identities, addressing the significant lack of diversity in literature. We present MIRRORSTORIES, a corpus of 1,500 personalized short stories generated by integrating elements such as name, gender, age, ethnicity, reader interest, and story moral. We demonstrate that LLMs can effectively incorporate diverse identity elements into narratives, with human evaluators identifying personalized elements in the stories with high accuracy. Through a comprehensive evaluation involving 26 diverse human judges, we compare the effectiveness of MIRRORSTORIES against generic narratives. We find that personalized LLM-generated stories not only outscore generic human-written and LLM-generated ones across all metrics of engagement (with average ratings of 4.22 versus 3.37 on a 5-point scale), but also achieve higher textual diversity while preserving the intended moral. We also provide analyses that include bias assessments and a study on the potential for integrating images into personalized stories.[1]

1 Introduction

"There is no greater agony than bearing an untold story inside you." — Maya Angelou

Mirror books are stories that reflect the reader's identity, culture, and experiences, serving to engage, validate, and empower individuals (Bishop, 1990). Such books are crucial in educational settings, fostering a sense of belonging and self-understanding through diverse narratives (Fleming et al., 2016), while also improving engagement and comprehension (Walkington and Bernacki, 2014; Heineke et al., 2022). Beyond education, personalized narratives have shown potential in various fields, including health communication and marketing, where they enhance patient understanding and adherence, and strengthen emotional connections between brands and consumers (Galitsky, 2024; Babatunde et al., 2024).

Despite the profound need for these personalized narratives, there is a noticeable underrepresentation of non-white minority groups in literature (CCBC, 2021) relative to their population size, detailed in Appendix Figure 6. The gap in cultural representation highlights the need for more inclusive narratives that reflect diverse reader identities, enhance empathy, and promote cultural awareness (Hoytt et al., 2022). Diversity in literature can lead to improved innovation and a broader consideration of ideas, ultimately enriching the reading experience for all (Phillips, 2014).

Advancements in natural language processing, particularly through the development of LLMs like GPT-4, PaLM, and LLaMA have introduced the potential to address these gaps on a large scale (OpenAI, 2023; Chowdhery et al., 2022; Touvron et al., 2023). LLMs excel in generating human-like text and adapting content to various contextual needs (Brown et al., 2020; Zhao et al., 2023).

Recent studies have investigated LLMs' capabilities in expressing personality within generated content (Li et al., 2024; Jiang et al., 2024) and developing methods to induce and edit personality expressions in LLM outputs (Jiang et al., 2023; Li et al., 2024; Mao et al., 2024). New benchmarks have also been released to assess personality traits in LLM outputs (Jiang et al., 2023; Wang et al., 2024). However, there remains a gap in research concerning whether LLMs can generate content that incorporates identity traits and faithfully mirrors the diverse identities of a global readership.

Our study addresses this gap by exploring the potential of LLMs to create mirror stories—narratives that genuinely reflect and resonate with the identities of individual readers. We present a framework

[1] Interactive web application and corpus are publicly available at mirrorstories.me.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6702–6717
November 12-16, 2024 ©2024 Association for Computational Linguistics

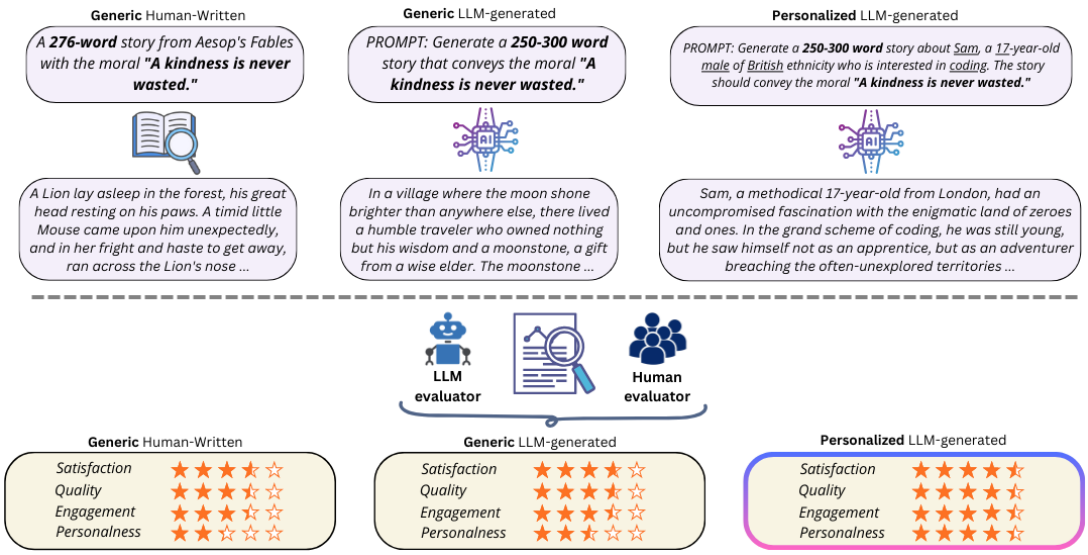


Figure 1: Generation and evaluation process for human-written generic, LLM-generated generic, and LLM-generated personalized narratives

that evaluates the effectiveness of LLM-generated mirror stories in comparison to traditional narratives, assessing their impact on engagement, satisfaction, and the perception of personal relevance (see Figure 1). Our contributions are three-fold:

1. We release MIRRORSTORIES, a corpus of 1,500 personalized short stories generated by integrating elements such as name, gender, age, ethnicity, reader interest, and moral of the story. We demonstrate that LLMs can effectively incorporate identity elements into narratives, with human evaluators identifying them in the stories with high accuracy.
2. Through a comprehensive evaluation involving 26 diverse human judges, personalized LLM-generated stories consistently outperform both generic human-written and LLM-generated stories across all engagement metrics, with a significantly higher average rating.
3. We present analyses that assess text diversity, coherence, and moral comprehension across each story type, and examine biases exhibited by LLMs when evaluating personalized narratives. We also explore the potential of integrating images and incorporate MIRRORSTORIES into an interactive **web application** where users can browse and generate stories.

2 MIRRORSTORIES

2.1 Overview

MIRRORSTORIES is a corpus designed to evaluate the ability of LLMs to generate both generic and personalized short narratives based on predefined morals and identity elements. Each dataset instance consists of a moral (e.g., "Kindness is never wasted") guiding the narrative's tone and a set of identity elements (name, age, gender, ethnicity, and personal interest) to personalize the story. Specifically, the dataset includes a human-written and an LLM-generated generic story, both of which do not incorporate specific identity elements, and an LLM-generated personalized story that includes these elements to enhance relevance and engagement. Appendix A.5 provides a detailed example of the dataset structure.

2.2 Dataset Collection

Human-written Stories & Morals MIRRORSTORIES incorporates human-written stories derived from Aesop's fables (Wier et al., 1890)[2], a well-known collection famous for its clear narrative structure and explicit moral conclusions. The morals serve as guides for generating both generic and personalized stories. The complete list of morals is provided in Appendix A.5 Table 7.

Identities Identity traits such as name, age, gender, ethnic background, and interests are included to personalize the narratives. We drew from 123 unique ethnic backgrounds, 124 diverse interests, and 28 distinct morals. The complete set of identities is provided in Appendix A.5 Table 7.

Generic & Personalized LLM-Generated Stories Generic stories are generated focusing solely

[2]Scraped from https://read.gov/aesop/001.html

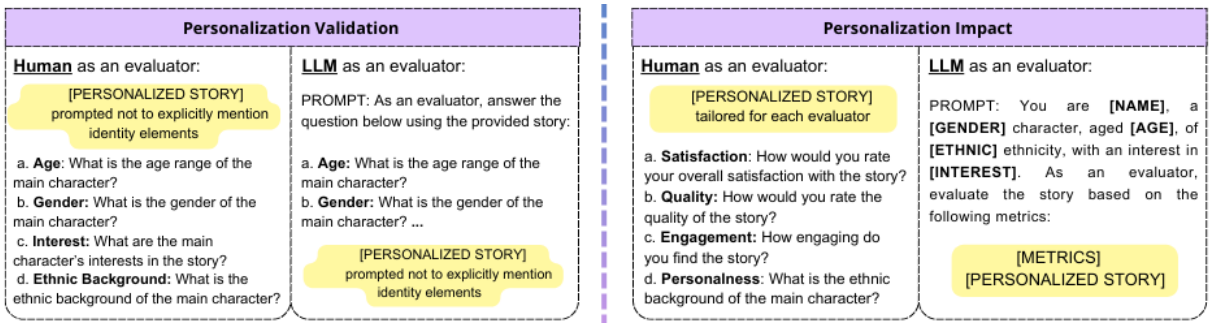


Figure 2: Illustration demonstrating the personalization validation and impact processes

on the moral, while personalized stories additionally integrate the specified identity elements. For the specific prompts, refer to Figure 1. GPT-4 (ver. 0613), Claude-3 Sonnet[3], and Gemini 1.5 Flash (Reid et al., 2024) were used, each responsible for generating one-third of the narratives.

MIRRORSTORIES comprises 1,500 narratives with an almost even split between male and female characters. The dataset spans a broad age range from 10 to 60 years. Detailed illustrations of the distributions are in Appendix A.1 Figures 9 and 10.

3 Experiments

We conducted two experiments to assess the effectiveness of personalization in LLM-generated stories: *Personalization Validation*, which validates the integration of identity elements within the narratives, and *Personalization Impact*, which assesses the impact of these narratives on user engagement, comprehension, satisfaction, and personalness.

Prompts In both experiments, personalized prompts incorporating identity elements were used to generate personalized stories. For Personalization Validation, these elements were specifically asked not to be stated explicitly, to test their seamless integration into the narrative. In the Personalization Impact experiment, personal elements were aligned with those of 26 human evaluators, ensuring that each story was tailored to evaluators. Figure 1 and 2 provide detailed structures of the prompts for both generation and evaluation.

Human Evaluation In both experiments, the same 26 human evaluators—all students from diverse disciplines—were tasked with evaluating the narratives (for demographic details, see Appendix Figure 8). For the Personalization Validation, they answered a structured questionnaire for a random sample of 30 stories to identify the personalized elements. In the Personalization Impact test, each evaluator reviewed a human-written, generic LLM-generated, and personalized LLM-generated story, with the personalized LLM-generated story specifically tailored to reflect their personal identity. They provided feedback on all three story types, rating them on satisfaction, quality, engagement, and personalness. The detailed questionnaire is provided in Appendix A.2.

Models GPT-4 (ver. 0613, temperature 0.4) was used as an evaluator in both experiments. Initially, it assessed the integration of personalized elements. Later, it was used to evaluate the stories for satisfaction, quality, engagement, and personalness, with a sample of the evaluation process and prompts provided in Appendix Figure 7. GPT-4 was chosen for its increasing adoption as an evaluator across domains (Gilardi et al., 2023; Tarkka et al., 2024; Malik et al., 2024), with potential advantages such as scalability, cost-efficiency, and consistency.

4 Results

Are MIRRORSTORIES personalized? The effectiveness of personalization in MIRRORSTORIES is evident from the high accuracy rates in identifying identity elements by both human and LLM evaluators. As shown in Figure 3, human evaluators were particularly adept at identifying gender and ethnicity with accuracies at 100% and 94%, respectively. Similarly, GPT-4 showed robust performance, matching or exceeding human accuracy in all categories, which confirms the high level of personalization achieved in the narratives.

Personalized LLM-generated stories also effectively incorporate both the provided moral and the reader's interests, with a stronger emphasis on the moral. To demonstrate this, we used BERTopic (Grootendorst, 2022) for topic modeling to identify the top five terms for each story. We then

[3] https://www.anthropic.com/claude

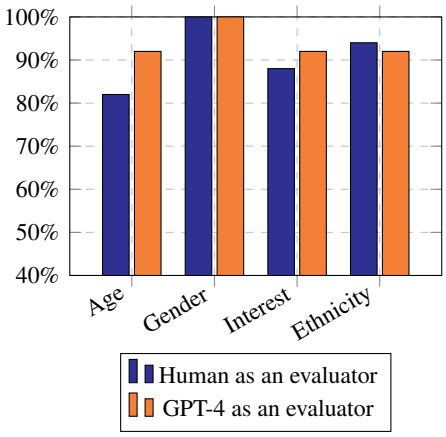


Figure 3: Accuracy of human and LLM evaluators in identifying identity elements in the story

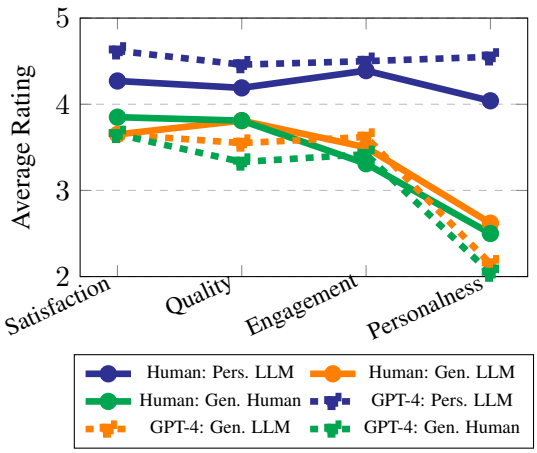


Figure 4: Comparative evaluation of narrative types by human and GPT-4 evaluators across different metrics

Story Type	Correctly Identified Morals
Generic Human-Written	23/24
Generic LLM-Generated	23/23
Personalized LLM-Generated	25/25

Table 1: Number of correctly identified morals for each story type, excluding 'N/A' responses

Story Type	SDI
Generic Human-Written	4.13
Generic LLM-Generated (temp = 1)	4.42
Generic LLM-Generated (temp = 1.2)	4.82[6]
Personalized LLM-Generated (all elements)	**4.71**
Personalized LLM-Generated (element: 'interest')	4.59

Table 2: The Shannon Diversity Index (SDI) values for all story types. Values are statistically significant ($p < 0.01$), as determined by a one-way ANOVA.

calculated cosine similarity using Word2Vec embeddings[4] (Mikolov et al., 2013) between these top terms and the provided interest and moral. The average cosine similarity was 0.12 for the provided interest and 0.27 for the moral, demonstrating a balance between incorporating the reader's interest and maintaining the intended moral.[5] A detailed sample of the top terms identified for each story is provided in Appendix Table 4.

Are MIRRORSTORIES preferred? Figure 4 shows that personalized LLM-generated stories in MIRRORSTORIES are consistently rated higher across all metrics compared to both generic LLM-generated and human-written narratives. This preference is pronounced in evaluations by both humans and GPT-4, with personalized narratives outperforming generic versions, particularly in terms of personalness and engagement where the ratings significantly diverge.

How does personalization affect moral comprehension? We analyzed the impact of personalization on moral comprehension in stories. Evaluators were asked to identify the main message of each type of story, or provide 'N/A' if they could not. We manually assessed the evaluators' responses to the intended morals. Excluding 'N/A' responses, the correctly identified morals are detailed in Table 1. The results indicate that differences in moral identification across story types are not statistically significant, demonstrating that adding personalization did not negatively affect the model's ability to convey the intended moral. A sample of evaluator responses is shown in Appendix Figure 3.

What is the impact of personalization on textual diversity? We analyzed how personalization elements impact textual diversity using the Shannon Diversity Index (SDI). Table 2 shows that personalized stories achieve the highest SDI among all story types. Including a single personalization element, such as the 'interest' element, also increases SDI compared to generic and human-written stories with the same moral. Additionally, we observed that increasing GPT-4's temperature negatively affects the diversity and coherence of generic LLM-generated stories. At a temperature of 1.2, the stories showed increased diversity but began to lose coherence. Further increasing the temperature to 1.5 resulted in nonsensical outputs.

[4] *word2vec-google-news-300*

[5] These cosine similarity values are relatively high; for context, the cosine similarity between the embeddings for *craft* and *carpentry* is 0.164.

[6] Increasing temperature made the stories lose coherence.

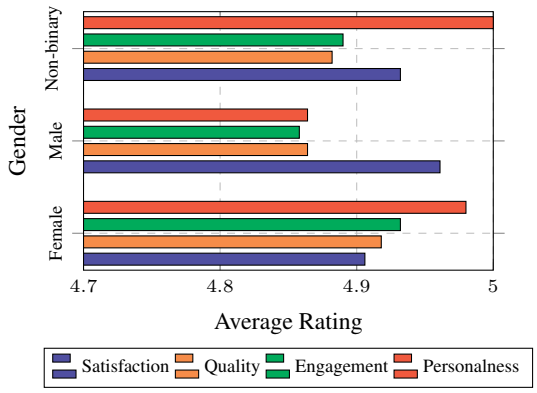


Figure 5: Average ratings by GPT-4 across gender

Are there biases in LLM evaluations of personalized stories? We found several preferential biases in GPT-4's evaluation results. Figure 5 shows an instance of gender-based bias, with stories featuring non-binary characters receiving the highest personalness ratings, while those with male characters rated lower in quality and engagement. Ethnic background also influences evaluations, with Norwegian and Japanese characters rated higher across all metrics (Appendix Figure 14). We also observed inter-model preferential biases across the three models used for generating personalized stories, with Claude-3 consistently receiving higher ratings compared to GPT-4 and Gemini-1.5. An overview of all bias results is provided in Appendix A.3.1.

5 Extended Analyses

Qualitative comparison of human and LLM evaluations We examine cases where human and LLM evaluators either contradict or agree on the scores assigned to stories, providing insights into the differences in evaluations and preferences for various types of stories. Examples of these cases, highlighting instances of both agreement and disagreement between human and LLM evaluators, are presented in Appendix Figure 15.

Image generation for personalized stories We explored the potential of incorporating images into stories to enhance engagement and representation. The image generation and evaluation processes are detailed in Appendix Table 18. Notably, human evaluators show a high accuracy in identifying personalized elements in the images generated by DALL·E 2 (Ramesh et al., 2022), with gender and interest being recognized with 100% and 95% accuracy, respectively (Appendix Figure 17).

Correlation between human and LLM evaluators Correlation analysis revealed a low to moderate alignment between human evaluators and GPT-4 in story evaluation metrics. GPT-4 aligned more closely with human evaluators on quality across all story types (correlations 0.22-0.47), but showed the weakest correlation in assessing personalness, particularly for personalized stories (as low as 0.08). This suggests that while GPT-4 is increasingly used for various evaluation tasks, its effectiveness in assessing subjective aspects of creative tasks is limited. A detailed analysis of these correlations and temperature variations is presented in Appendix A.4.2, Table 5 and Figure 16.

6 Related Work

Our study builds on research on the effectiveness of personalized narratives in engaging readers and improving learning outcomes (Zhang et al., 2024; Pennebaker and Graybeal, 2001; Hirsh and Peterson, 2009). We extend this work by examining how LLMs can generate personalized narratives to increase reader engagement and satisfaction. While promising, the accuracy of personal traits in generated content remains challenging, with studies showing mixed results (Jiang et al., 2024; Bhandari and Brennan, 2023). Concurrently, LLM exploration in narrative generation has focused on improving coherence and depth (Andreas, 2022; Shen and Elhoseiny, 2023; El-Refai et al., 2024; Gómez-Rodríguez and Williams, 2023). To assess these advancements, recent evaluative techniques for narrative systems emphasize user interactions and alignment metrics between visual content and narratives (El-Refai et al., 2024; Ning et al., 2023).

7 Conclusion

Our study demonstrates the potential of LLMs in generating personalized narratives that effectively incorporate diverse identity elements and enhance reader engagement compared to generic stories. MIRRORSTORIES consists of 1,500 personalized stories that consistently outperform generic ones on key metrics. By making MIRRORSTORIES publicly available and integrating it into an interactive web application, we aim to encourage further research on personalized narrative generation, contributing to more engaging and inclusive content. Future work could explore out-group perceptions of these narratives, broadening our understanding of personalization's impact across diverse audiences.

Limitations

Story Constraints: To maintain consistency and feasibility within the scope of our study, we imposed certain constraints on the stories generated, such as limiting the length to 250-300 words and focusing on a specific set of morals. While these constraints allowed for a controlled comparison between personalized and generic stories, they may not fully capture the potential of LLMs in generating longer, more complex narratives or exploring a wider range of themes and morals. Future research could investigate the impact of personalization on stories of varying lengths and themes to gain a more comprehensive understanding of how these factors influence reader engagement and satisfaction.

Demographic Diversity: While our study aimed to include a diverse range of identities and backgrounds, the demographic diversity of our human evaluators was by no means the perfect sample of global readership. The majority of our evaluators were university students, which may not be representative of the broader population. Future research should include a more diverse pool of evaluators across age, education, and cultural backgrounds to ensure the generalizability of the findings and to capture a wider range of perspectives on personalized storytelling.

Scope of Personalization: Our study primarily examined personalization factors like age, gender, interests, and ethnic background. However, aspects such as personality traits, emotional resonance, and narrative preferences were not extensively investigated but could notably enhance engagement and narrative impact. For example, aligning story elements with reader emotional responses or tailoring narratives to specific preferences like mystery, romance, or adventure could significantly boost satisfaction and engagement.

Subjectivity of Evaluation: Another limitation of our study is the inherent subjectivity involved in evaluating the impact of personalized stories. Despite our attempts to standardize evaluation criteria and maintain consistency among evaluators, individual preferences, biases, and interpretations can still significantly influence the outcomes. This subjectivity can lead to variability in how different evaluators perceive and rate the same narrative elements.

Model Selection and Variety: Our study utilized GPT-4, Claude3, Gemini-1.5, and DALL·E 2 for generating and evaluating narratives and images. This limited selection may affect the generalizability of our findings, as different models might produce or assess stories differently based on their training data and algorithms. Expanding future research to include a variety of models, including open-source ones, could provide a more comprehensive understanding of how different language models handle personalization in storytelling and evaluate narrative elements.

Ethical Considerations

We followed strict ethical standards throughout our research to ensure validity and fairness. Consent and transparency were central to our approach, with all participants fully informed and providing explicit consent. We also ensured compliance with intellectual property rights by using Aesop's fables, which are in the public domain.

Data Privacy and Security: Ensuring the privacy and security of participants' personal information was a top priority. We collected and used personal details such as age, gender, interests, and ethnic background to generate personalized stories. Robust data protection measures were implemented, including secure storage, anonymization, and restricted access to sensitive information. Participants were informed about how their data would be used, stored, and protected.

Potential Misuse and Unintended Consequences: While personalized storytelling has the potential to enhance engagement and representation, we carefully considered the potential for misuse or unintended consequences. To mitigate risks such as the manipulation of individuals' emotions or the reinforcement of stereotypes, we implemented safeguards against harmful content and regularly audited the generated stories for potential biases or inappropriate themes.

Inclusivity and Representation: When generating personalized stories, we strived to ensure that the stories were inclusive and representative of diverse identities and experiences. This included considering factors such as race, ethnicity, gender identity, sexual orientation, disability, and socioeconomic status. We aimed to create stories that were respectful, authentic, and empowering for all individuals, avoiding stereotypes and promoting positive representation.

Accountability and integrity were paramount in

reporting our results, including limitations and implications. Additionally, every narrative generated by LLMs underwent a thorough review to maintain quality and appropriateness, enhancing the reliability of our findings and participant well-being.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada and by the New Frontiers in Research Fund.

References

Jacob Andreas. 2022. Language models as agent models.

Sodiq Babatunde, Opeyemi Odejide, Tolulope Edunjobi, and Damilola Ogundipe. 2024. The role of ai in marketing personalization: A theoretical exploration of consumer engagement strategies. *International Journal of Management & Entrepreneurship Research*, 6:936–949.

Prabin Bhandari and Hannah Marie Brennan. 2023. Trustworthiness of children stories generated by large language models. *arXiv preprint arXiv:2308.00073*.

Rudine Sims Bishop. 1990. Mirrors, windows, and sliding glass doors. perspectives: Choosing and using books for the classroom, 6 (3). *Perspectives: Choosing and using books for the classroom*, 6(3):ix–xi.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

CCBC. 2021. Books by and/or about black, indigenous, and people of color (all years). Data retrieved from the Cooperative Children's Book Center.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Karim El-Refai, Zeeshan Patel, and Jonathan Pei. 2024. Swag: Storytelling with action guidance. *arXiv preprint arXiv:2402.03483*.

Jane Fleming, Susan Catapano, Candace M Thompson, and Sandy Ruvalcaba Carrillo. 2016. *More mirrors in the classroom: Using urban children's literature to increase literacy*. Rowman & Littlefield.

Boris A. Galitsky. 2024. Llm- based personalized recommendations in health. *Preprints*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of llms on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, A Coruña, Spain and Sunshine Coast, Australia. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Amy J. Heineke, Aimee Papola-Ellis, and Joseph Elliott. 2022. Using texts as mirrors: The power of readers seeing themselves. *The Reading Teacher*, 76(3):277–284.

Jacob Hirsh and Jordan Peterson. 2009. Personality and language use in self-narratives. *Journal of Research in Personality*, 43:524–527.

Karima Hoytt, Sherrica Hunt, and Margaret A Lovett. 2022. Impact of cultural responsiveness on student achievement in secondary schools. *Alabama Journal of Educational Leadership*, 9:1–12.

D Huyck and SP Dahlen. 2019. Diversity in children's books 2018. sarahpark. com blog. created in consultation with edith campbell, molly beth griffin, kt horning, debbie reese, ebony elizabeth thomas, and madeline tyner, with statistics compiled by the cooperative children's book center, school of education, university of wisconsin-madison.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. Personallm: Investigating the ability of large language models to express personality traits.

Tianlong Li, Shihan Dou, Changze Lv, Wenhao Liu, Jianhan Xu, Muling Wu, Zixuan Ling, Xiaoqing Zheng, and Xuanjing Huang. 2024. Tailoring personality traits in large language models via unsupervisedly-built personalized lexicons.

Usman Malik, Simon Bernard, Alexandre Pauchet, Clément Chatelain, Romain Picot-Clémente, and Jérôme Cortinovis. 2024. Pseudo-labeling with large language models for multi-label emotion classification of french tweets. *IEEE Access*, 12:15902–15916.

Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. 2024. Editing personality for large language models.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Munan Ning, Yujia Xie, Dongdong Chen, Zeyin Song, Lu Yuan, Yonghong Tian, and Qixiang Ye. 2023. Album storytelling with iterative story-aware captioning and large language models. *arXiv preprint arXiv:2305.12943*.

OpenAI. 2023. Gpt-4 technical report.

James W. Pennebaker and Anna Graybeal. 2001. Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science*, 10(3):90–93.

Katherine Phillips. 2014. How diversity works. *Scientific American*, 311:42–7.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Xiaoqian Shen and Mohamed Elhoseiny. 2023. Storygpt-v: Large language models as consistent story visualizers. *arXiv preprint arXiv:2402.03483*.

Otto Tarkka, Jaakko Koljonen, Markus Korhonen, Juuso Laine, Kristian Martiskainen, Kimmo Elo, and Veronika Laippala. 2024. Automated emotion annotation of Finnish parliamentary speeches using GPT-4. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 70–76, Torino, Italia. ELRA and ICCL.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Candace Walkington and Matthew Bernacki. 2014. *Motivating Students by "Personalizing" Learning around Individual Interests: A Consideration of Theory, Design, and Implementation Issues*, volume 18, chapter 4. Preprints.

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews.

H. Wier, J. Tenniel, and E.H. Griset. 1890. *Aesop's Fables: A New Revised Version from Original Sources*. Worthington, Company.

Chao Zhang, Xuechen Liu, Katherine Ziska, Soobin Jeon, Chi-Lin Yu, and Ying Xu. 2024. Mathemyths: Leveraging large language models to teach mathematical language through child-ai co-creative storytelling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–23, New York, NY, USA. ACM.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

A Appendix

Figure 6: Illustration of diversity representation in children's books based on 2018 publishing statistics. Data derived from the Cooperative Children's Book Center, University of Wisconsin-Madison. (Illustration by David Huyck, in consultation with Sarah Park Dahlen) (Huyck and Dahlen, 2019).

GPT-4 as an Evaluator

Sample GPT-4 INPUT:
You are ANDRII, a MALE character, aged 25, of RUSSIAN ethnicity, with an interest in WRESTLING. As an evaluator, evaluate the story based on the following metrics:

METRICS:

SATISFACTION: How would you rate your overall satisfaction with the story?
1 (BAD): The story did not meet any expectations, had significant flaws in plot, character development, or other key areas, and provided little to no enjoyment or value.
5 (GOOD): The story exceeded all expectations, was exceptional in most if not all aspects.
QUALITY: How would you rate the quality of the story? ...

STORY:
In the heart of Moscow lived Andrii, a 25-year-old whose heart and soul were intertwined with the art of wrestling. Every morning, before the first ray of sunlight kissed the domes of Saint Basil's Cathedral ...

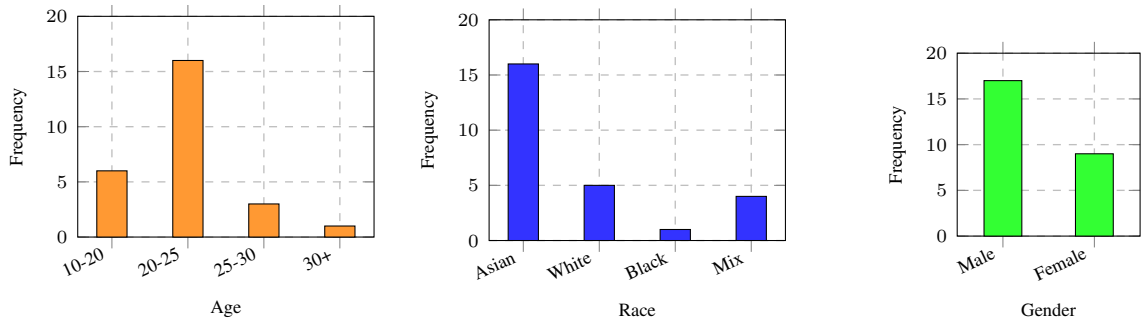
Sample GPT-4 OUTPUT:
I would rate my overall satisfaction with the story a 5. The unexpected turn in the plot, well-developed character, and the consistent twists kept me engaged, and the story exceeded all my expectations.
I would rate the quality of content of the story a 4. The intertwining of fiction and reality, the character ...

Figure 7: Sample input and output of GPT-4 evaluating a personalized story

Annotator #	Story Type	Q: **"What is the main message or moral of the story? If you can't identify, write 'N/A' ''.**	Actual Moral of the Story
Annotator #1	Generic Human-Written	To preserver and never quit. The tortoise never stopped during the race but the hare did and ended up losing because of it.	The race is not always to the swift.
Annotator #1	Generic LLM-Generated	Perseverance. Keep your eye on the prize, don't worry about what others are doing.	The race is not always to the swift.
Annotator #1	Personalized LLM-Generated	Perseverance. Not everyone's goals in life will be the same and will take time. Keep taking steps towards your goal and you will be rewarded.	The race is not always to the swift.
Annotator #2	Generic Human-Written	Despite one's status someone can still help another.	Kindness is never wasted.
Annotator #2	Generic LLM-Generated	DIn order to truly help somebody, one should do their best to accommodate their needs and avoid demanding some kind of compensation for their troubles.	Kindness is never wasted.
Annotator #2	Personalized LLM-Generated	That kindness has a way of repaying itself one day.	Kindness is never wasted.

Table 3: Sample responses from two annotators on the main message or moral of each story type, compared with the actual intended moral of the stories

A.1 Annotators and Dataset Diversity



(a) **Age Distribution** (b) **Race Distribution** (c) **Gender Distribution**

Figure 8: Demographic distribution of annotators by age, race, and gender

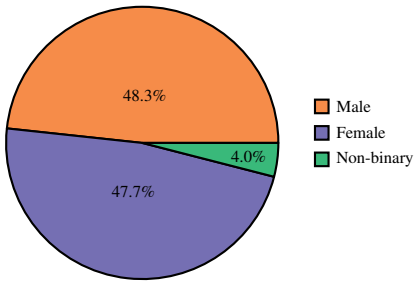


Figure 9: Gender Distribution in MIRRORSTORIES

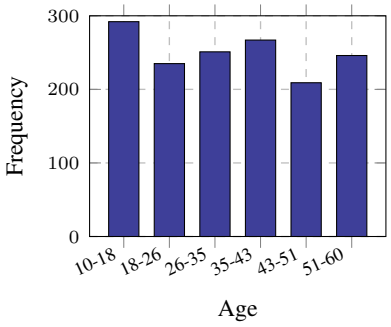


Figure 10: Age Distribution in MIRRORSTORIES

A.2 Questionnaire for Annotators

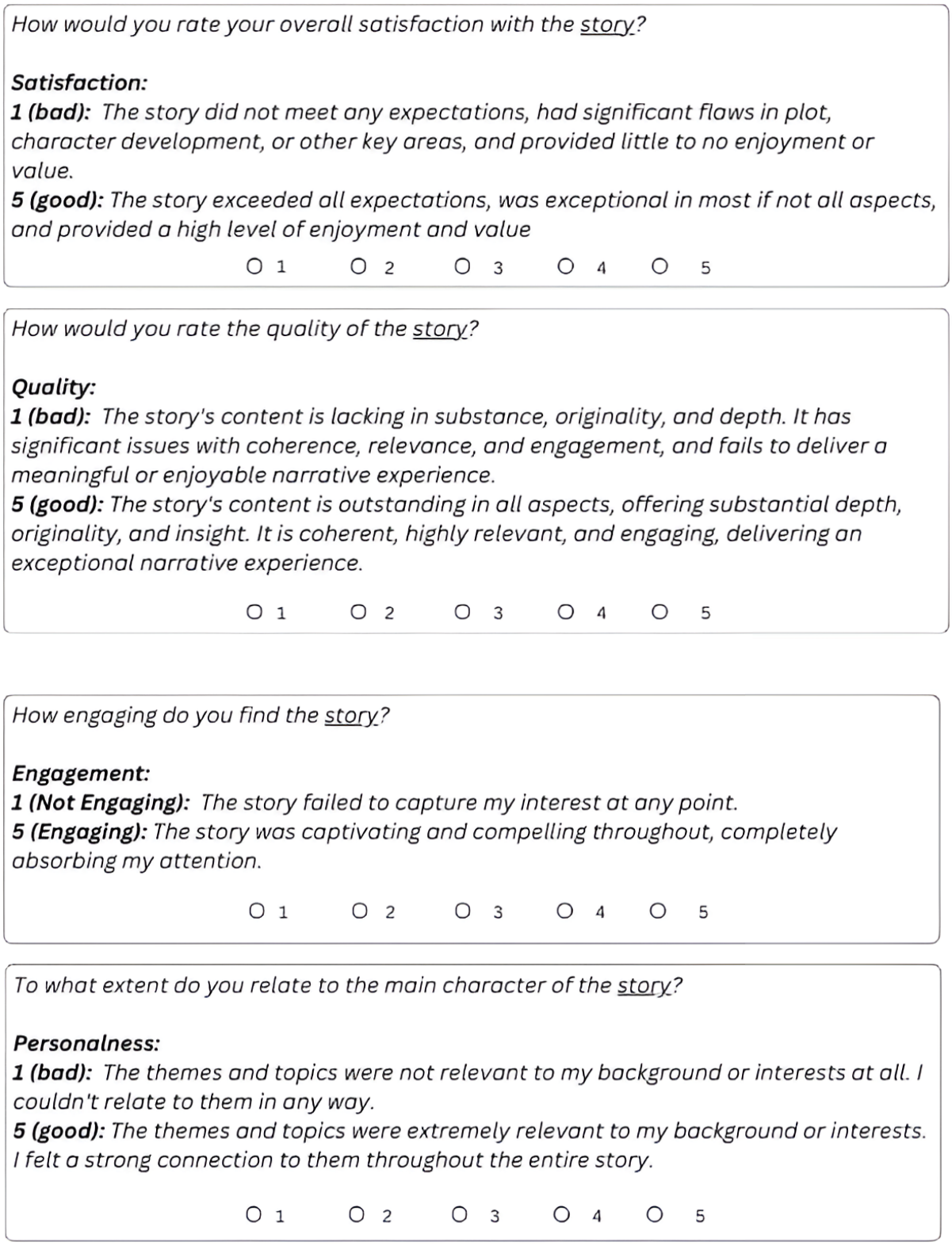


Figure 11: Questionnaire used to assess story satisfaction, quality, engagement, and personalness

A.3 Personalization Example

Personalization Elements	Personalized Story	Top 5 Term
Aveline, 19, Non-binary, Reading, French	Aveline was a confluence of distinctive characteristics; their name a symbol of French roots and their gender identity, non-binary, a testament to their unfettered self-expression. At 19, they were a sagacious soul, finding immense joy in ...	library (0.033), truth (0.030), guilt (0.023), joy (0.023), lie (0.023)
Farida, 23, Female, Carpentry, Uzbek	In the heart of the bustling Uzbek city, Tashkent, Farida, a young, passionate woman, was determined to carve out a unique reputation for herself. At 23, she was an anomaly in her city; she was not working a typical job like ...	craft (0.0397), reputation (0.039), wood (0.039), client (0.029), city (0.029)
Rami, 21, Male, Trekking, Syrian	Rami, a 21-year-old Syrian youth, was known in his community for two things — his irresistible passion for trekking and his firm belief in honesty. Dark-haired, with ...	community (0.030), truth (0.028), life (0.022), sun (0.022), adventurous (0.021)

Table 4: This table presents the personalization elements for three individuals, their personalized stories, and the top 5 terms identified by BERTopic for each story, along with the corresponding relevance scores.

A.3.1 Preferential Bias Analysis

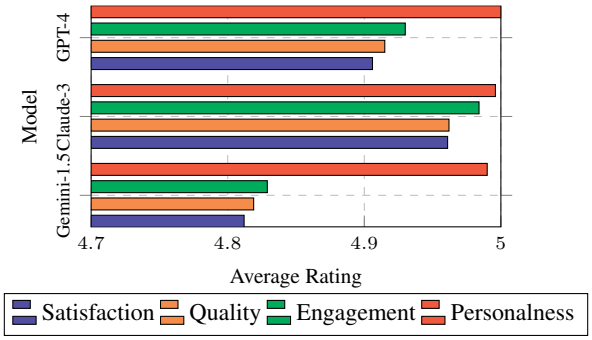


Figure 12: Average evaluation ratings by GPT-4 across models

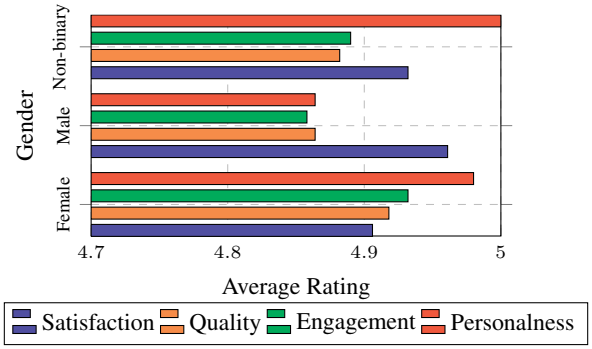


Figure 13: Average evaluation ratings by GPT-4 across gender

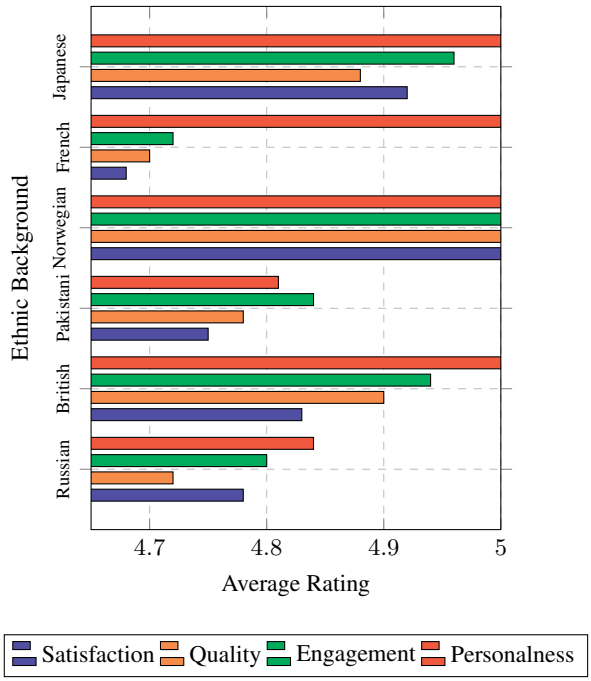


Figure 14: Average evaluation ratings by GPT-4 across ethnic background

A.4 Extended Analysis

A.4.1 Qualitative Analysis

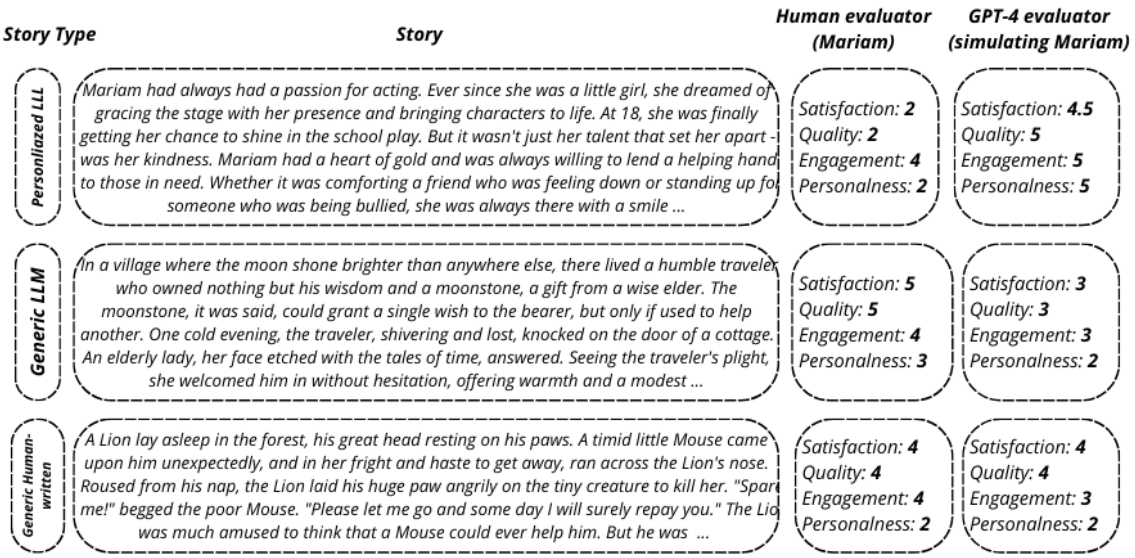


Figure 15: Qualitative comparison of ratings for three types of stories by both human evaluators and GPT-4, including both conflicting and consistent ratings

A.4.2 Correlation Analysis

Story Type	Satisfaction	Quality	Engagement	Personalness
Personalized LLM-generated	0.24	0.47	0.34	0.08
Generic LLM-generated	0.19	0.22	0.20	0.12
Generic Human-written	0.12	0.26	0.21	0.19

Table 5: Spearman's rank correlation coefficients between human evaluators and GPT-4 for story evaluation metrics, where values closer to 1 indicate a stronger positive correlation

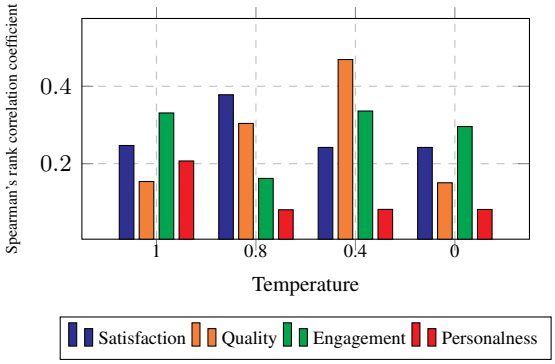


Figure 16: Spearman's rank correlation coefficient between human evaluators and GPT-4 at various temperatures for personalized LLM-generated stories

A.4.3 Image generation for personalized stories

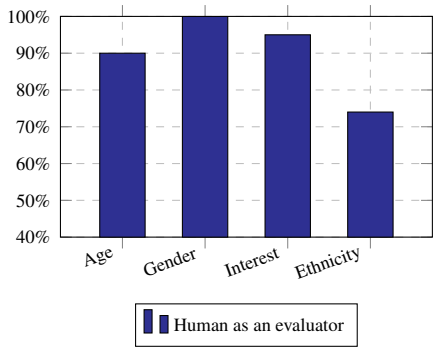


Figure 17: Accuracy of human and LLM evaluators in identifying personalization elements in the image

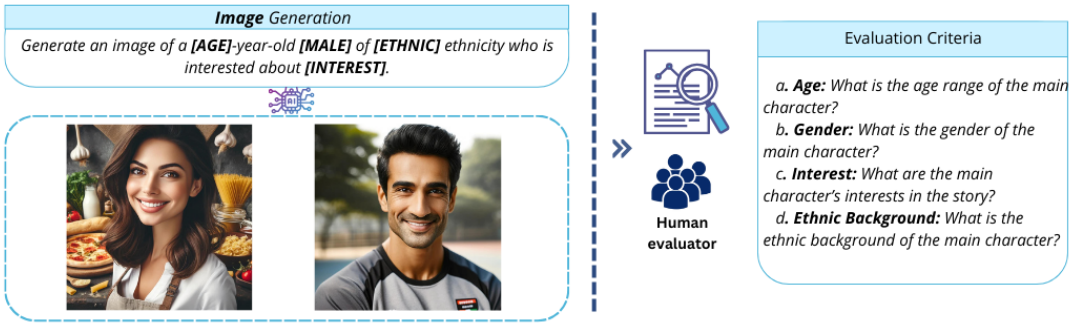


Figure 18: Illustration of the personalization test process for images. The left side displays the prompt used to generate personalized images. The right side outlines the evaluation criteria for human evaluators to determine how effectively personal elements have been integrated into the image.

A.5 Dataset Structure, Categories and Values

Name	Age	Gender	Interest	Ethnicity	Moral	LLM-generated Personalized Story	LLM-generated Generic Story	Human-written Generic Story
Behzad	23	Male	Boxing	Indian	Learn from your mistakes	Behzad, a sturdy 23-year-old Indian man, was far more attached to his boxing gloves than his engineering textbooks. Son of a mechanic, his life was perched comfortably on the rusty edges of Mumbai, lit up by the lambent moon by night and the intruding sun by day. Unlike his peers, Behzad wasn't fascinated by the jingling call of corporate ladders; instead, he ...	...	...
Karina	29	Female	History	Ukrainian	Strive for excellence	Karina's eyes danced with curiosity as she traced her fingers over the ancient tome. The worn leather binding and yellowed pages spoke of centuries past, each word a whisper from a forgotten era. For her, the study of history was more than a mere pursuit of knowledge, it was a gateway to understanding ...	...	...
Dario	35	Non-binary	Volunteering	Italian	Take responsibility for your actions	Dario's eyes sparkled with determination as they stepped into the volunteer center. At 35 years old, they had already made a name for themselves in the local community for their unwavering commitment to making a difference. Born to Italian parents who instilled a deep sense of responsibility and compassion ...	...	...

Table 6: Dataset structure of MIRRORSTORIES. The dataset includes personal attributes (Name, Age, Gender, Interest, Ethnicity), a moral, and three types of stories: LLM-Generated Personalized Story, LLM-Generated Generic Story, and Human-Written Generic Story.

Category	Values
Age	10 ... 60.
Gender	Male, Female, Non-binary.
Ethnicity	Albanian, Arab, Arab-Amazigh, Armenian, Australian, Austrian, Akan, Andorran, Azerbaijani, Bambara, Belarusian, Bengali, Baganda, Bosnian, Pardo Brazilian, Bosniak, British, Bulgarian, Canadian, Chechen, Chinese, Congolese, Croat, Czech, Dane, Dutch, Egyptian, Emirati, Estonian, Fijian, Finn, Georgian, German, Greek, Hawaiian, Hungarian, Indian, Indonesian (Javanese), Iraqi Arab, Irish, Italian, Japanese, Jewish, Jordanian, Kazakh, Korean, Kikuyu, Kurdish, Kuwaiti, Kyrgyz, Latvian, Lithuanian, Luxembourger, Malay, Maldivian, Maltese, Maori, Mestizo, Moldovan, Norwegian, Punjabi, Palestinian, Persian, Polish, Portuguese, Romanian, Russian, Hutu, Salvadoran, Scottish, Serb, Slovak, Slovene, Somali, Spanish, Sudanese, Swede, Swiss, Syrian, Tajik, Thai, Turk, Turkmen Ukrainian, Uzbek, Vietnamese (Viet), Welsh, Wolof.
Interest	Acting, Archery, Arts, Astronomy, Badminton, Bagpiping, Baking, Ballet, Baseball, Basketball, Beadwork, Beekeeping, Biology, Biking, Blogging, Board, Bonsai, Boxing, Calligraphy, Camping, Canoeing, Carpentry, Chess, Coding, Community Service, Cooking, Crafting, Cricket, Culinary, Cycling, Dancing, Digital, Drawing, Drumming, Embroidery, Falconry, Farming, Fashion, Fencing, Filmmaking, Fishing, Football, Foraging, Gardening, Geography, Graphic design, Guitar, Gymnastics, Hiking, History, Hockey, Horseback, Ice, Ikebana, Judo, K-Pop, Kayaking, Kendo, Kickboxing, Kite Flying, Knitting, Literature, Jewelry Making, Martial Arts, Massage, Meditation, Mountaineering, Music, Painting, Papercraft, Parkour, Photography, Piano, Pilates, Poetry, Politics,; Pottery, Quilting, Reading, Respect, Riding, Robotics, Rock Climbing, Rowing, Rugby, Running, Sailing, Sculpting, Science, Sewing, Skateboarding, Skydiving, Skiing, Singing, Social Activities, Soccer, Sprinting, Storytelling, Surfing, Swimming, Taekwondo, Table, Tango, Teaching, ennis, Traveling, Trekking, Video Games, Violin, Volleyball, Volunteering, Weaving, Weightlifting, Winemaking, Woodworking, Wrestling, Writing, Yoga.
Moral	Maintain humility, Learn from your mistakes, Be optimistic, Show empathy, Be loyal, Work hard and stay humble, Live with purpose, Take responsibility for your actions, Always tell the truth, Cherish your family, Be generous, Keep your promises, Treat others as you want to be treated, Be a good listener, Be fair and just, Live with integrity, Protect the weak and vulnerable, Seek justice, Be curious and keep learning, Be grateful, Have courage, Help those in need, Strive for excellence, Have respect for yourself and others, Practice good manners, Embrace diversity, The race is not always to the swift, A kindness is never wasted, Liars are not believed even when they speak the truth.

Table 7: Breakdown of the different categories and values included in the MIRRORSTORIES dataset. It covers a diverse range of ages, genders, ethnicities, interests, and moral values.