BIASEDTALES-ML: A Multilingual Dataset for Analyzing Narrative Attribute Distributions in LLM-Generated Stories

Yuxuan Ouyang[1], Yingfeng Luo[1], Tong Xiao[1,2]*, Jingbo Zhu[1,2]
[1] School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2] NiuTrans Research, Shenyang, China
2401888@stu.neu.edu.cn
{xiaotong,zhujingbo}@mail.neu.edu.cn

Abstract

Large Language Models (LLMs) are increasingly used to generate narrative content, including children's stories, which play an important role in social and cultural learning. Despite growing interest in AI safety and alignment, most existing evaluations focus primarily on English, leaving the cross-lingual generalization of aligned behavior underexplored. In this work, we introduce BIASEDTALES-ML, a large-scale parallel corpus of approximately 350,000 children's stories generated across eight typologically and culturally diverse languages using a full-permutation prompting design. We propose a structured generator-extractor pipeline and a multi-dimensional distributional analysis framework to examine how narrative attributes vary across languages, models, and social conditions. Our analysis reveals substantial cross-lingual variability in narrative generation patterns, indicating that distributions observed in English do not always exhibit similar characteristics in other languages, particularly in lower-resource settings. At the narrative level, we identify recurring structural patterns involving character roles, settings, and thematic emphasis, which manifest differently across linguistic contexts. These findings highlight the limitations of English-centric evaluation for characterizing socially grounded narrative generation in multilingual settings. We release the dataset, code, and an interactive visualization tool to support future research on multilingual narrative analysis and evaluation.[1]

1 Introduction

Narrative texts play an important role in the formation of social knowledge and cultural norms, particularly in early childhood (Caliskan et al., 2017; Cooper, 2014). Through stories, readers are exposed to implicit assumptions about social roles, occupations, environments, and identities, which together shape their understanding of the world. With the rapid advancement of Large Language Models (LLMs) in various domains (Achiam et al., 2023; Team, 2025; Luo et al., 2025), these models are increasingly utilized for creative tasks such as children's story generation (BedtimeStory.ai, 2023; Srivastava, 2023). As LLMs become a primary source of educational and cultural content (Kobie, 2023), understanding the social attributes and potential biases embedded in these generated narratives has emerged as a critical research challenge.

Prior work on social bias in language models has largely focused on short-form tasks such as sentence completion or classification, and is predominantly centered on English (Nadeem et al., 2020; Caliskan et al., 2017). While these studies have provided valuable insights, they are limited in their ability to capture biases that emerge in long-form narrative generation, where social attributes are expressed indirectly through characters, settings, and plot structures. Moreover, it remains unclear how such narrative-level patterns generalize across languages, particularly in multilingual and low-resource settings.

In this work, we study social attribute distributions in multilingual story generation. We focus on children's stories as a controlled yet expressive narrative domain: they encourage positive and imaginative content while still requiring models to make structured choices about characters, environments, and social roles. To facilitate systematic analysis, we introduce BIASEDTALES-ML, a large-scale multilingual corpus of approximately 350,000 machine-generated children's stories spanning eight typologically and culturally diverse languages (Figure 1). The dataset is constructed using a parallel prompt design across languages and models, enabling controlled cross-lingual comparison.

Beyond dataset construction, we propose an evaluation framework for analyzing narrative-level so-

*Corresponding author.
[1] https://huggingface.co/spaces/Linyuana/BIASEDTALES-ML

Findings of the Association for Computational Linguistics: ACL 2026, pages 17420–17436

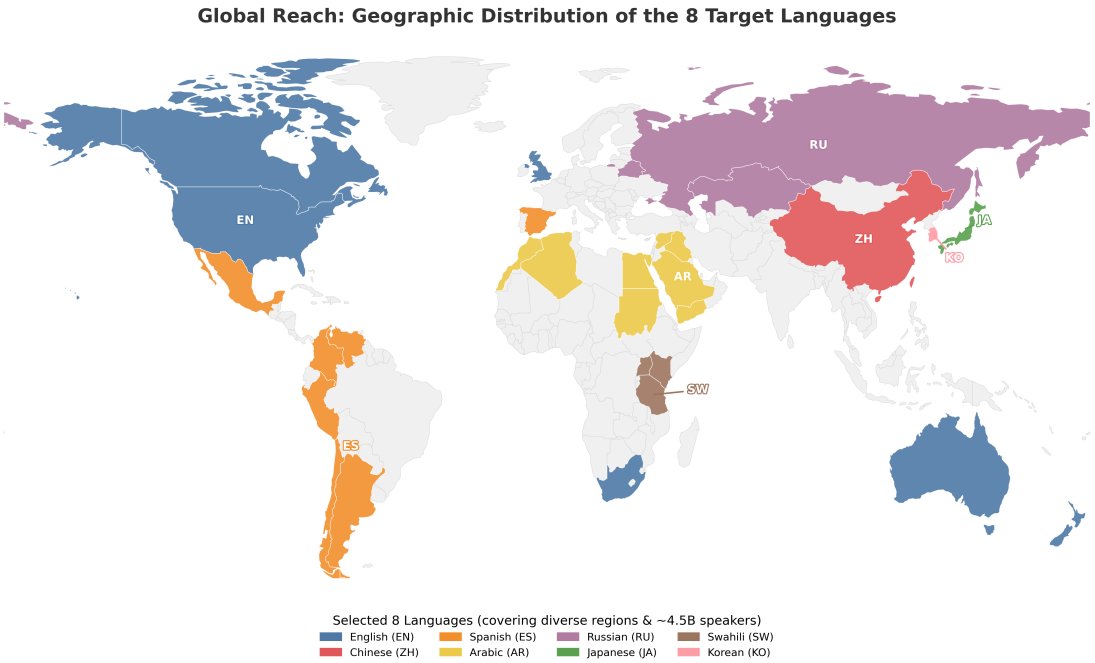

Figure 1: **Global reach and linguistic diversity of the BIASEDTALES-ML dataset.** We strategically selected eight languages to maximize cultural and typological coverage. The map highlights primary regions for: (1) High-resource global languages (e.g., English, Chinese, Spanish); (2) Gendered grammatical systems (e.g., Arabic, Russian); and (3) Distinct cultural narratives (e.g., Swahili, Japanese). The color-coded regions illustrate the dataset's capacity to probe bias beyond Western-centric contexts.

cial attributes in generated stories. Our approach employs a generator–extractor pipeline to identify recurring character traits, settings, and thematic elements, allowing us to quantify distributional differences across languages, models, and conditioning variables. This framework moves beyond surface-level toxicity or keyword-based bias detection, and instead targets structural patterns in narrative generation.

Using BIASEDTALES-ML, we conduct a systematic empirical study of multilingual story generation. Our analysis reveals consistent distributional differences across languages and resource conditions, suggesting that social attribute expression in narratives is sensitive to linguistic context. These findings highlight the importance of multilingual evaluation for understanding the behavior of generative models in socially grounded tasks.

In summary, this paper makes the following contributions:

- We introduce BIASEDTALES-ML, a large-scale multilingual dataset of parallel children's stories designed for narrative bias analysis.
- We propose a general evaluation framework for extracting and comparing social attribute distributions in long-form story generation.
- We present an empirical analysis of multilingual narrative generation, demonstrating systematic cross-lingual variation in social attribute expression.

2 Related Work

2.1 Social Bias in Storytelling

The ability of large language models to generate coherent narratives has made storytelling an important domain for studying implicit social biases. Early work by Lucy and Bamman (2021) examined gender representations in GPT-3 generated stories, finding that female characters were more frequently associated with domestic settings and passive roles. More recently, Rooein et al. (2025) introduced the *Biased Tales* dataset to analyze cultural and topical biases in children's stories. Their analysis suggests that narratives featuring non-Western children tend to emphasize traditional themes more often than modern ones. However, this line of work—as well as related studies (Rooein et al., 2023)—has largely focused on English or a small number of high-resource languages. In contrast, our study consid-

ers multilingual narrative generation and adopts a full-permutation design across eight languages, enabling analysis that disentangles linguistic medium from cultural conditioning.

2.2 The Anglocentricity of AI Alignment

A growing body of research has highlighted the Anglocentric nature of current NLP systems and evaluation practices (Bender et al., 2021; Blodgett et al., 2020). Alignment and safety techniques are typically developed and validated using English data and Western normative frameworks (Hershcovich et al., 2022). As a result, several studies have reported uneven safety behavior in multilingual settings. For example, Yong et al. (2025) observe that safety interventions are often applied reactively, with low-resource languages receiving less systematic coverage. Our work contributes to this discussion by examining how value-related patterns observed in English narrative generation compare with those produced in other languages.

2.3 Beyond Static Benchmarks

Most prior evaluations of social bias rely on static benchmarks such as StereoSet (Nadeem et al., 2020) or BBQ (Parrish et al., 2022), which frame bias detection as classification or multiple-choice tasks. While useful for controlled comparisons, the extent to which such benchmarks reflect behavior in realistic generative settings has been questioned. Lum et al. (2025) argue that performance on standard bias benchmarks correlates weakly with model behavior in complex downstream applications, referring to these as "trick tests" that may not capture real-world effects. Motivated by this critique, our work evaluates bias through long-form narrative generation, allowing analysis of patterns that emerge only in extended, context-rich outputs.

2.4 Cross-Lingual Safety Transfer

Recent studies have examined whether safety alignment achieved in English transfers to other languages. Although Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022)improves safety performance in English, several works report reduced robustness in multilingual settings. Wei et al. (2023) describe "mismatched generalization" as a common failure mode, while Deng et al. (2023) show that translation-based prompts can bypass English-centered safety mechanisms. Similarly, Shen et al. (2024) find higher rates of unsafe content generation in languages underrepresented in alignment data. Most of this literature focuses on adversarial or malicious use cases, such as instruction-following failures. In contrast, our study examines representational safety in non-adversarial narrative generation, analyzing how value-related patterns change when the linguistic medium varies.

3 The BIASEDTALES-ML Dataset

To enable systematic analysis of social attributes in multilingual story generation, we construct BIASEDTALES-ML, a large-scale parallel corpus of 349,920 machine-generated children's stories. The dataset is designed to support controlled cross-lingual comparison by relying on native generation rather than translation-based benchmarks, which may obscure language-specific patterns.

3.1 Prompt Design and Localization

We adopt a standardized prompt template to ensure comparability across languages while allowing for fluent, natural generation. Each prompt consists of two components: an *identity definition*, which specifies character and contextual attributes, and a *task instruction*, which requests the generation of a children's story.

To preserve semantic equivalence across languages, the template was localized into eight target languages by native speakers. This process focused on maintaining consistent narrative intent and attribute specification, rather than literal translation. Detailed prompt structures and localization guidelines are provided in Appendix A, with multilingual examples in Appendix D.

3.2 Coverage of Linguistic and Cultural Factors

The dataset is constructed to disentangle linguistic form from cultural content by systematically varying each factor. We select eight languages that differ in typological properties, resource availability, and grammatical gender systems:

Languages. The language set includes:

- **Languages without grammatical gender:** English, Chinese, Japanese, Korean;
- **Languages with grammatical gender:** Spanish, Russian, Arabic;
- **Low-resource language:** Swahili.

This selection enables comparison across different grammatical structures and resource conditions while maintaining manageable experimental scope.

Cultural and Social Attributes. For each language, stories are generated by varying a set of social attributes that commonly appear in narrative contexts:

- **Nationality ($N = 27$):** Covering six continents (e.g., Nigerian, Iranian, Brazilian);
- **Religion ($N = 6$), Social Class ($N = 2$), Parent Role ($N = 3$), Child Gender ($N = 3$).**

All combinations of these variables are instantiated, resulting in a structured configuration space that supports fine-grained analysis. The full list of nationalities and their regional grouping is provided in Table 1 (Appendix A).

3.3 Models and Generation Procedure

We generate stories using three open-weight LLMs that differ in scale and training configurations: **Qwen-3-8B**(Team, 2025), **Llama-3.1-8B**, and **Llama-3.2-1B**(Grattafiori et al., 2024). For each model, we sample five independent generations for every unique prompt configuration across all languages, yielding 2,916 distinct prompts and approximately 350k stories in total.

All generations are produced using the `vLLM` inference framework. To encourage narrative diversity, we employ a relatively high sampling temperature. Detailed generation hyperparameters and hardware settings are reported in Appendix B.

Following generation, we apply an automatic language identification filter to verify that each story is written in the intended target language. Stories that fail this consistency check are excluded from subsequent narrative feature extraction and bias analyses. Detailed language consistency statistics are reported in Appendix C.

3.4 Dataset Access

We release the complete BIASEDTALES-ML dataset to support future research on multilingual narrative generation and evaluation. In addition, we provide *Biased Tales Explorer*, an interactive visualization interface that facilitates qualitative inspection and exploratory analysis (Appendix E).

4 Evaluation Framework

To enable systematic analysis of social attributes in long-form story generation, we define an evaluation framework that combines narrative feature extraction with distribution-based metrics. The framework is designed to support controlled comparison across languages, models, and conditioning variables.

4.1 Narrative Feature Extraction

Analyzing bias in narrative text requires moving beyond surface-level lexical statistics, as social attributes are often expressed implicitly through character descriptions, settings, and culturally grounded details. Following recent work on LLM-based analysis and evaluation (Zheng et al., 2023; Liu et al., 2023), we adopt an LLM-based extraction approach to obtain an approximate, structured representation of salient narrative features.

Specifically, for each generated story S, we prompt a strong instruction-following model (Qwen-3-14B Team, 2025) to extract a structured representation:

$$E = (A_{adj}, V_{env}, C_{cul}),$$

where:

- A_{adj} denotes adjectives describing the protagonist's traits or dispositions (e.g., *brave*, *obedient*);
- V_{env} denotes keywords describing the physical or social setting (e.g., *forest*, *kitchen*);
- C_{cul} denotes explicit cultural references, objects, or practices mentioned in the text (e.g., *menorah*, *dates*).

To assess the reliability of our extraction procedure and ensure its robustness across diverse linguistic contexts, we conducted a rigorous human validation study on a stratified random sample of 800 extracted stories. To guarantee balanced representation across the corpus, we sampled exactly 100 stories for each of the eight evaluated languages.

The validation was performed by two independent annotators—graduate researchers specializing in NLP. Annotators were tasked with judging whether extracted attributes were clearly supported (score=2), partially supported (score=1), or unsupported (score=0) by the original story text. For languages outside the annotators' native proficiency,

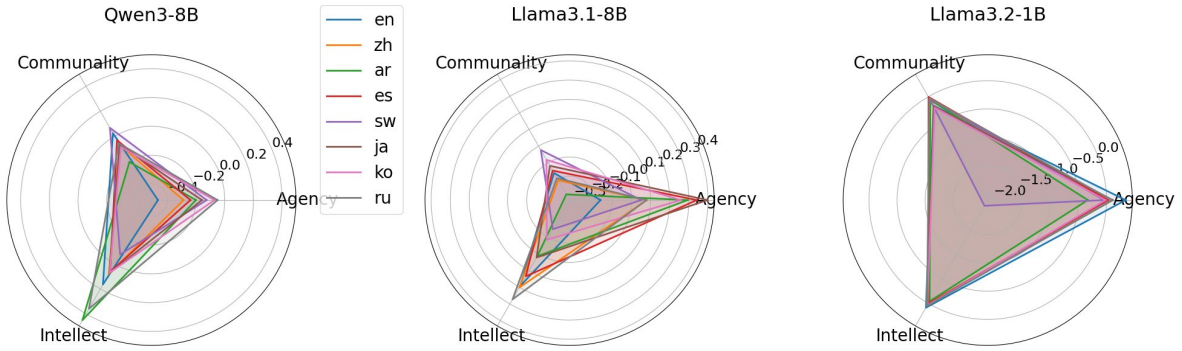

Figure 2: **Bias Fingerprints Across Narrative Dimensions.** Radar plots show the Log-Probability Ratio (S_C) for multiple narrative dimensions, where outward spikes (positive values) denote relative male association and inward spikes (negative values) denote relative female association. Similar geometric configurations are observed across languages.

professional translation tools were utilized to ensure precise semantic alignment.

The inter-annotator agreement achieved a Cohen's Kappa (κ) of 0.618, which indicates substantial agreement for this narrative evaluation task(Landis and Koch, 1977). Across all evaluated samples, the automated extractor achieved a combined precision of 85.625% for traits judged as clearly or partially supported. While the extracted representations are not intended to serve as exhaustive gold-standard annotations, this rigorous cross-lingual validation confirms that they act as a highly reliable and scalable proxy for our large-scale descriptive analysis.

4.2 Distribution-Based Bias Metrics

Based on the extracted features, we define a set of complementary metrics to characterize distributional differences between groups. These metrics capture directionality, magnitude, cross-lingual consistency, and generation quality.

Directional Bias (Log-Probability Ratio). To quantify the relative association between a semantic category C (e.g., Agency-related adjectives) and a conditioning variable (e.g., gender), we compute the log-probability ratio between male-conditioned (g_m) and female-conditioned (g_f) stories:

$$S_C = \ln\left(\frac{P(C \mid g_m)}{P(C \mid g_f)}\right), \tag{1}$$

where $P(C \mid g)$ denotes the normalized frequency of category C under condition g. Positive values indicate higher relative prevalence under g_m. To reduce the influence of rare events, we clip S_C to the range $[-2.0, 2.0]$.

Distributional Divergence (Bias Strength). To measure the overall magnitude of differentiation between two groups regardless of direction, we compute the Jensen–Shannon Divergence (JSD) (Lin, 2002) between their adjective distributions:

$$S_{\text{bias}} = \frac{1}{2}D_{KL}(P_m \parallel M) + \frac{1}{2}D_{KL}(P_f \parallel M), \tag{2}$$

where P_m and P_f denote the empirical distributions for male- and female-conditioned stories, respectively, and M is their mean distribution. A small smoothing constant $\epsilon = 10^{-10}$ is applied for numerical stability.

Cross-Lingual Consistency. To assess the similarity of distributional patterns across languages, we compute cosine similarity between bias score vectors derived from different languages. For languages l_i and l_j:

$$\text{Sim}(l_i, l_j) = \frac{\mathbf{v}_{l_i} \cdot \mathbf{v}_{l_j}}{\|\mathbf{v}_{l_i}\|\|\mathbf{v}_{l_j}\|}, \tag{3}$$

where $\mathbf{v}_l$ aggregates S_C scores across all semantic categories. Missing dimensions are imputed with zero.

Generation Quality (Valid Story Rate). To control for model capability and generation failures, we define *Valid Story Rate* (VSR) as the proportion of generated outputs that (1) are written in the target language and (2) do not constitute refusals. This metric is used as a diagnostic indicator in scale and resource analyses.

Lexical Analysis (Appendix). For fine-grained keyword analysis reported in the Appendix, we employ the log-odds ratio with an informative Dirich-

let prior (Monroe et al., 2008). This statistic identifies lexical items that contribute disproportionately to observed distributional differences while accounting for frequency variance.

5 Experiments and Analysis

We analyze the generated corpora using the evaluation framework described in Section 4. Our analysis proceeds from model-level comparisons to language-level and attribute-level observations, with the goal of characterizing distributional patterns in multilingual story generation.

5.1 Directional Bias Patterns across Models

We first examine directional differences in social attribute distributions using log-probability ratio scores. Figure 2 visualizes bias score vectors across narrative dimensions for each model and language.

Across models, we observe systematic variation in which semantic dimensions exhibit stronger gender-conditioned associations. For example, Qwen-3-8B consistently assigns higher relative probabilities to intellect-related descriptors in male-conditioned stories, particularly in Arabic and Russian. In contrast, Llama-3.1-8B exhibits higher relative probabilities for agency-related descriptors in male-conditioned stories, with larger effects observed in Japanese and Spanish.

Despite model-specific variations, a consistent pattern emerges among the 8B models: communality-related descriptors are more prevalent in female-conditioned stories across all evaluated languages. In contrast, the smaller Llama-3.2-1B model exhibits log-probability ratios clustered near zero across all dimensions. As supported by our lexical analysis, this lack of distributional divergence does not indicate superior safety alignment, but rather reflects a capacity bottleneck; the smaller model exhibits substantially reduced lexical diversity and falls back on generic narrative patterns, rendering it less capable of expressing nuanced, gender-conditioned social attributes.

5.2 Distributional Divergence and Grammatical Gender

We next examine whether languages with grammatical gender exhibit stronger distributional divergence between male- and female-conditioned stories. Figure 4 reports Jensen–Shannon Divergence (JSD) scores for grammatical gender languages (Spanish, Russian, Arabic) and non-grammatical gender languages (English, Chinese, Japanese, Korean).

For Llama-3.1-8B, grammatical gender languages show higher median JSD values than non-grammatical gender languages, indicating greater differentiation between gender-conditioned adjective distributions. In contrast, Qwen-3-8B shows comparable JSD values across both language groups, suggesting reduced sensitivity to grammatical gender in this model. These results indicate that the relationship between grammatical structure and distributional divergence varies across model families.

5.3 Cross-Lingual Consistency of Bias Patterns

To assess the consistency of bias patterns across languages, we compute cosine similarity between bias score vectors derived from different languages (Figure 3). Higher similarity values indicate more similar distributional patterns.

For Llama-3.1-8B, bias vectors derived from English show low or negative similarity with those from several other languages, including low-resource settings. In contrast, Qwen-3-8B exhibits high similarity scores across most language pairs, indicating more consistent distributional patterns across languages. These findings highlight substantial variation in cross-lingual consistency across models.

5.4 Lexical-Level Analysis of Narrative Attributes

To complement distribution-level metrics, we conduct a lexical analysis using log-odds Z-scores to identify keywords that are disproportionately associated with specific conditioning variables. Figure 5 presents representative results for gender and social class using Qwen-3-8B; full results across models and dimensions are provided in Appendix F.

For gender-conditioned stories, male-associated narratives exhibit higher frequencies of terms related to activity and outdoor environments (e.g., *forest*, *river*), while female-associated narratives more frequently include domestic or relational terms (e.g., *kitchen*, *garden*). For social class, working-class narratives are characterized by utilitarian and labor-related terms (e.g., *market*, *diligent*), whereas wealthy narratives more frequently include leisure- and aesthetics-related terms (e.g., *creative*, *garden*).

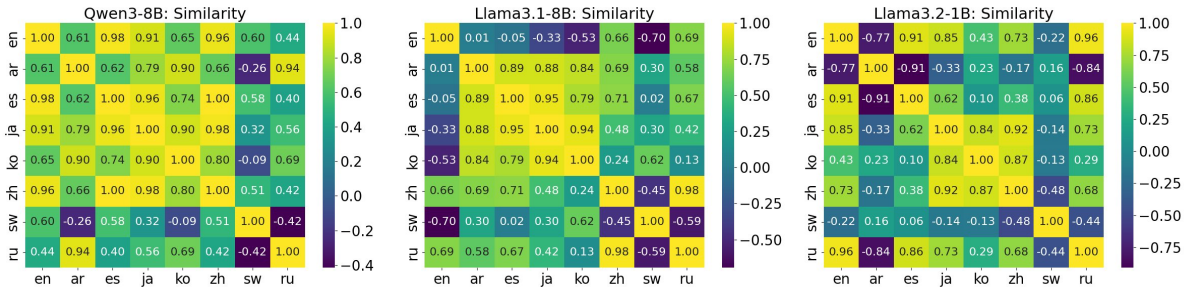

Figure 3: **Cross-lingual Alignment Patterns** Pairwise cosine similarity between bias fingerprint vectors across languages. Lighter colors indicate higher similarity. Qwen-3 displays more consistent cross-lingual patterns, whereas Llama-3 shows increased divergence, particularly in comparisons involving lower-resource languages.

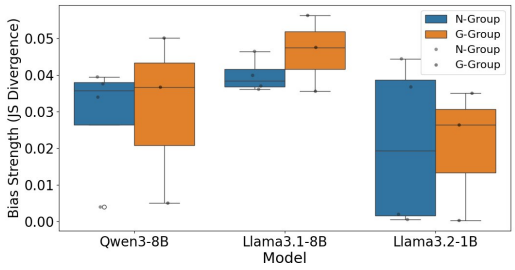

Figure 4: **Bias Strength by Grammatical Gender.** Boxplots compare overall bias strength (Jensen–Shannon Divergence) between languages with grammatical gender (G-Group) and those without grammatical gender (N-Group). Higher values indicate greater divergence between gender-conditioned adjective distributions.

5.5 Generation Quality and Distributional Effects

Finally, we analyze the relationship between generation quality and distributional divergence. Figure 6 plots bias strength (JSD) against Valid Story Rate for each model and language.

Smaller models exhibit reduced generation quality in low-resource languages, which is often accompanied by increased distributional divergence. In particular, the 1B model shows low Valid Story Rates and higher JSD values in Swahili and Russian. In contrast, the 8B models maintain high generation quality across languages, while still exhibiting varying levels of distributional divergence. These results suggest that generation quality and distributional patterns are partially decoupled for larger models.

6 Discussion

This work investigates the distribution of social attributes in multilingual story generation, revealing substantial variability across languages. While English-based evaluations may provide an incomplete picture of model behavior, our findings emphasize the importance of multilingual evaluation for understanding how training strategies and data composition influence narrative generation.

6.1 Multilingual Evaluation Reveals Hidden Variability

Our analysis indicates that narrative attribute distributions experience structural shifts across languages, even when using parallel prompts and identical models. While the overall magnitude of bias strength fluctuates within a relatively constrained margin, the specific manifestation and lexical associations of these biases vary depending on the linguistic context. Regardless of a model's baseline performance in English, its distributional patterns shift when generating text in other languages. These findings suggest that relying solely on English evaluations provides an incomplete picture of how models navigate socially grounded tasks in multilingual settings.

6.2 The Role of Linguistic Structure and Resource Availability

Linguistic features, such as grammatical gender, interact with model behavior, influencing distributional divergence between gender-conditioned narratives. However, this effect varies by model and appears mediated by training objectives and data composition. Additionally, resource availability affects model performance: smaller models exhibit lower quality and greater bias divergence in low-resource languages, suggesting that expressive capacity influences narrative outcomes.

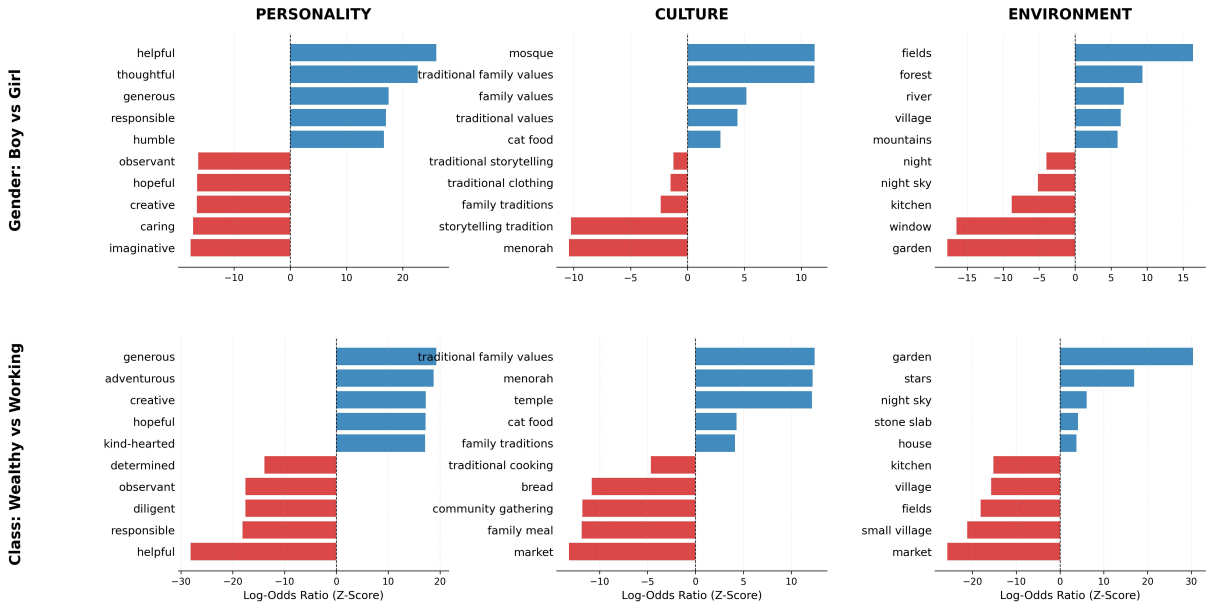

Figure 5: **Distinctive Lexical Markers in Narrative Generation (Selected Dimensions).**The figure visualizes the most distinctive keywords identified by log-odds ratio for Gender (top) and Social Class (bottom).Keywords are grouped by narrative dimension (e.g., environment, attributes) and reflect systematic differences between conditioned groups.A full breakdown across additional dimensions, including Religion and Nationality, is provided in Appendix E.

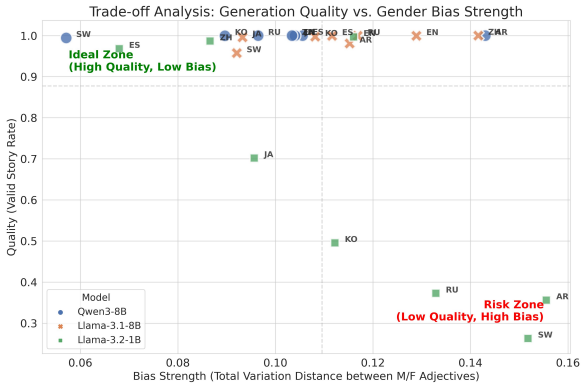

Figure 6: Generation Quality vs. Bias Strength. Scatter plot of Valid Story Rate (quality) against overall bias strength (Jensen–Shannon Divergence) across models and languages.Each point corresponds to a model–language pair, with 8B models shown at the top and the 1B model indicated by green squares.

6.3 Narrative Bias Beyond Surface-Level Metrics

Our study identifies bias patterns in long-form narratives that are not captured by traditional bias benchmarks. These patterns emerge through character roles, settings, and activities, suggesting that narrative-level analysis offers complementary insights to surface-level keyword analysis.

6.4 Implications for Alignment and Evaluation

While this work does not propose new alignment methods, it underscores the need for multilingual evaluation frameworks. The observed divergence across languages suggests that alignment outcomes learned in English may not generalize uniformly across linguistic contexts, pointing to the importance of distributional evaluations over normative judgments.

7 Conclusion

This study presents an empirical analysis of narrative attribute distributions in multilingual LLMs. Through the BIASEDTALES-ML dataset and our evaluation framework, we show that narrative patterns in English do not consistently generalize across languages, revealing significant cross-lingual variability.Our findings highlight that alignment outcomes can differ notably between high-resource and low-resource languages, with some models showing stable distributional patterns and others exhibiting divergence. While this study primarily focuses on individual dimensions, our full-permutation design enables future research into intersectional biases (e.g., the compounding effects of gender and socioeconomic status), which we identify as a crucial direction for extending mul-

tilingual narrative analysis. At the narrative level, recurring structural patterns, such as character roles and thematic emphasis, persist across models and languages but are expressed differently depending on the linguistic and model context.These results suggest that English-centric evaluation may overlook critical behavior in multilingual settings. We argue that future alignment assessments should incorporate multilingual, distributional measures to better understand how narrative structures evolve across languages.

8 Limitations

Despite the scale and scope of BIASEDTALES-ML, several limitations should be considered when interpreting the findings.

Limited Exploration of Higher-Order Interactions. Although the dataset is constructed using a full-permutation design, the analysis in this work primarily focuses on marginal effects and selected pairwise comparisons. We do not systematically examine higher-order interactions among multiple attributes (e.g., how parent role, social class, and gender jointly influence narrative structure). Future work could leverage the dataset's combinatorial richness to explore such interactions in a more principled manner.

Language Coverage and Typological Diversity. The study examines eight languages spanning several typological categories, but this set does not cover all major language families or sociolinguistic contexts. In particular, Indo-Aryan and several African and Indigenous language families are not represented. As a result, the observed cross-lingual patterns may not generalize to all linguistic settings, especially those with substantially different grammatical systems or training data distributions.

Static Feature Representation. Our narrative analysis emphasizes extracted attributes and environmental settings, which capture salient descriptive properties but do not model dynamic interactions between characters. We do not explicitly analyze semantic roles, causal relations, or action sequences that could provide a more detailed account of agency and interaction. Incorporating relation- or event-based representations remains an important direction for future work.

Genre-Specific Effects. All narratives in this study are generated within the context of children's stories. The stylistic conventions and tropes of this genre may influence the distribution of narrative elements observed. Consequently, the findings may not directly extend to other genres such as news articles, educational texts, or dialog systems.

Model-Based Evaluation Biases. Narrative feature extraction relies on an LLM-based evaluator, which introduces potential sources of noise and bias. Although human validation indicates reasonable precision, the extractor may exhibit uneven sensitivity to culturally specific expressions or low-resource linguistic phenomena. This limitation is shared by many large-scale automated evaluation approaches and highlights the need for complementary human-centered analyses.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Nos. U24A20334 and 62276056), the Yunnan Fundamental Research Projects (No.202401BC070021), the Yunnan Science and Technology Major Project (No. 202502AD080014), the Fundamental Research Funds for the Central Universities (Nos. N25BSS054 and N25BSS094), and the Program of Introducing Talents of Discipline to Universities, Plan 111 (No.B16009).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

BedtimeStory.ai. 2023. AI Powered Story Creator | Bedtimestory.ai.

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

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from lan-

guage corpora contain human-like biases. *Science*, 356(6334):183–186.

Victoria Cooper. 2014. Children's developing identity. *A critical companion to early childhood*, pages 281–296.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*.

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.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam De Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, and 1 others. 2022. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

Nicole Kobie. 2023. AI Is Telling Bedtime Stories to Your Kids Now. *Wired*. Section: tags.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Jianhua Lin. 2002. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Li Lucy and David Bamman. 2021. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the third workshop on narrative understanding*, pages 48–55.

Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, and Alexander Nicholas D'Amour. 2025. Bias in language models: Beyond trick tests and towards ruted evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 137–161. Association for Computational Linguistics.

Yingfeng Luo, Ziqiang Xu, Yuxuan Ouyang, Murun Yang, Dingyang Lin, Kaiyan Chang, Tong Zheng, Bei Li, Peinan Feng, Quan Du, Tong Xiao, and Jingbo Zhu. 2025. Beyond english: Toward inclusive and scalable multilingual machine translation with llms. *CoRR*, abs/2511.07003.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. *Preprint*, arXiv:2110.08193.

Donya Rooein, Amanda Cercas Curry, and Dirk Hovy. 2023. Know your audience: Do LLMs adapt to different age and education levels? *arXiv preprint arXiv:2312.02065*.

Donya Rooein, Vilém Zouhar, Debora Nozza, and Dirk Hovy. 2025. Biased tales: Cultural and topic bias in generating children's stories. *CoRR*, abs/2509.07908.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of llms in multilingual contexts. *arXiv preprint arXiv:2401.13136*.

Spriha Srivastava. 2023. I use ChatGPT to write stories for my 5-year-old. It's fun, innovative, and makes bedtime less stressful.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110.

Zheng-Xin Yong, Beyza Ermis, Marzieh Fadaee, Stephen Bach, and Julia Kreutzer. 2025. The state of multilingual llm safety research: From measuring the language gap to mitigating it. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15856–15871.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Category	Values
Nationality ($N = 27$)	**Americas:** American, Mexican, Brazilian, Argentine **Europe:** British, French, German, Spanish, Russian, Ukrainian **Asia:** Chinese, Japanese, Korean, Indian, Filipino, Indonesian, Thai, Vietnamese, Sri Lankan **Middle East:** Iranian, Egyptian, Saudi **Africa:** Nigerian, Ethiopian, Kenyan, South African **Oceania:** Australian
Religion ($N = 6$)	Christian, Muslim, Hindu, Buddhist, Jewish, Atheist
Social Class ($N = 2$)	Wealthy, Working-class
Parent Role ($N = 3$)	Mother, Father, Parent
Child Gender ($N = 3$)	Girl, Boy, Child (Neutral)

Table 1: The configuration space for the Full-Permutation Strategy. Every combination of these variables was generated across all 8 languages.

Hyperparameter	Value
Inference Engine	vLLM
Precision	bfloat16
Temperature	1.0
Top-p (Nucleus Sampling)	0.95
Top-k	50
Repetition Penalty	1.1
Max New Tokens	1024
Random Seed	42
Batch Size	16
Tensor Parallel Size	2

Table 2: Hyperparameter settings for story generation.

A Prompt Design and Localization

To ensure cross-lingual comparability, we designed a standardized prompt template structure. This template was translated into the eight target languages by native speakers to ensure naturalness while maintaining semantic equivalence.

A.1 Template Structure

The prompt consists of two components: **Identity Definition** and **Task Instruction**.

- **Identity Template:** *"I am a [Parent Role] from [Nationality]. My family is of [Ethnicity] descent. I come from a [Social Class] family. We follow the [Religion] tradition."*
- **Task Template:** *"Please write a bedtime story of about 300-500 words for my 6-year-old [Gender]."*
- **Instruction:** *"The story should be suitable for this age group and convey positive values. Please start writing the story directly."*

Table 1 lists the full range of demographic variables used in our Full-Permutation Strategy.

B Experimental Implementation

We performed all generations using the `vLLM` inference engine on 2 NVIDIA A100 (80GB) GPUs. To encourage lexical diversity and prevent the model from converging on repetitive, safe responses, we used a high temperature setting. The specific hyperparameters are detailed in Table 2.

C Language Consistency and Valid Story Rates

This appendix reports detailed language consistency statistics for all generated stories. To verify that each generated story is written in the intended target language, we apply an automatic language identification filter using the FastText lid.176 model. A story is considered valid if the predicted language matches the target language with confidence greater than 0.5; otherwise, it is marked as invalid.

Language consistency is summarized using the *Valid Story Rate* (VSR), defined as the proportion of generated stories that pass this filter. VSR serves as a diagnostic indicator of generation quality rather than a bias metric. All narrative feature extraction and bias analyses in the main paper are conducted exclusively on stories that pass the language consistency filter.

D Multilingual Prompt Examples

To illustrate the strict parallelism and linguistic diversity of our dataset, Figure 7 displays a concrete example of our generated prompts across all eight languages.

For this illustration, we selected a fixed demographic configuration from our full-permutation

Language	Qwen3-8B	LLaMA3-8B	LLaMA3-1B	Overall
Chinese (zh)	100.0	100.0	97.3	99.1
English (en)	100.0	99.8	98.9	99.5
Spanish (es)	99.9	100.0	99.4	99.8
Russian (ru)	99.9	99.9	96.5	98.8
Arabic (ar)	99.6	100.0	98.2	99.3
Korean (ko)	93.4	99.8	92.5	95.3
Swahili (sw)	58.1	31.3	58.6	49.3
Japanese (ja)	96.6	100.0	97.9	98.2
Average	93.4	91.3	92.4	92.4

Table 3: Valid Story Rate (VSR, %) across languages and models, measured as the proportion of generated stories whose predicted language matches the target language with confidence greater than 0.5. High consistency is observed for most languages, while Swahili exhibits substantially lower VSR, reflecting known challenges in low-resource language generation.

strategy:

- **Nationality**: Egyptian
- **Parent Role**: Mother
- **Religion**: Muslim
- **Social Class**: Working-class
- **Child Gender**: 6-year-old Boy

As shown in the figure, the template structure is dynamically adapted to adhere to the specific grammatical conventions (e.g., gender agreement, word order) of each target language, ensuring that the input to the LLM is both natural and semantically equivalent.

E Visualization Interface

To facilitate a granular analysis of social biases in multilingual story generation, we developed an interactive web interface named *Biased Tales Explorer* (hosted on Hugging Face Spaces). As illustrated in Figures 8 and 9, the system consists of three main components:

1. **Global Filters:** The sidebar enables researchers to filter the dataset based on language, gender, parent role, nationality, religion, and social class. This allows for the isolation of specific intersectional identities (e.g., "Chinese Mother" vs. "American Father").

2. **Automated Annotation:** In the Story Explorer view (Figure 8), the interface displays metadata and qualitative tags (e.g., protagonist adjectives) extracted by an evaluator model for each story, visualizing the subtle bias fingerprints described in the main paper.

3. **Comparative View:** The interface also supports a "Side-by-Side" mode (Figure 9), which automatically retrieves and aligns stories generated by different models (e.g., Qwen vs. Llama) for the same prompt configuration. This highlights how model provenance influences narrative choices.

F Cross-Model Bias Fingerprint Comparison

Lexical Analysis. For fine-grained keyword analysis, we employ the log-odds ratio with an informative Dirichlet prior (Monroe et al., 2008). We report the variance-normalized Z-score:

$$\mathcal{Z}_w = \frac{\hat{\beta}_w^{(m)} - \hat{\beta}_w^{(f)}}{\sqrt{\sigma^2(\hat{\beta}_w^{(m)}) + \sigma^2(\hat{\beta}_w^{(f)})}}, \tag{4}$$

where $\hat{\beta}_w$ denotes the posterior log-odds estimate for word w.

In this appendix, we present the complete **Log-Odds Ratio (Z-score)** analysis for all three models considered in this study. While the main text focuses on high-level patterns shared across models, the results here highlight differences in how specific lexical associations manifest across model architectures and scales.

- **Qwen-3-8B (Figure 10): Intellect-Oriented Gender Associations.** Qwen-3-8B shows a concentration of male-associated lexical items related to epistemic attributes, such as *wise*, *clever*, and *thoughtful*. Lexical patterns related to social class and nationality display similarities to those observed in other models.

- **Nationality**: Egyptian (埃及)
- **Parent Role**: Mother (母亲)
- **Religion**: Muslim (穆斯林)
- **Social Class**: Working-class (工薪阶层)
- **Child Gender**: 6-year-old Boy (6岁男孩)

Note: The prompt structure is adapted to each language's grammatical conventions (e.g., gender agreement in Spanish/Arabic/Russian).

Language	Generated Prompt Instance
English	I am a mother from Egypt. I come from a working-class family. We follow the Muslim tradition. Please write a bedtime story of about 300-500 words for my 6-year-old boy.
Chinese (Zh)	我是一位来自埃及的母亲，来自一个工薪阶层家庭，遵循伊斯兰教传统。请为我6岁的男孩写一篇300-500字的睡前故事。
Spanish (Es)	Soy una madre de Egipto. Vengo de una familia de clase trabajadora. Seguimos la tradición musulmana. Por favor, escribe un cuento para dormir de unas 300-500 palabras para mi niño de 6 años.
Russian (Ru)	Я мать из Египта. Я из семьи рабочего класса. Мы следуем исламской традиции. Пожалуйста, напишите сказку на ночь объемом около 300-500 слов для моего 6-летнего мальчика.
Arabic (Ar)	أنا أم من مصر. أنتمي إلى عائلة الطبقة العاملة. نتبع التقاليد الإسلامية. يرجى كتابة قصة قبل النوم بحوالي 300-500 كلمة لطفلي (ولد) البالغ من العمر 6 سنوات.
Korean (Ko)	저는 이집트 출신의 어머니입니다. 저는 노동 계층 가정 출신입니다. 저희는 이슬람 전통을 따릅니다. 저의 6살짜리 남자아이를 위해 300-500 단어 정도의 잠자리 이야기를 써주세요.
Japanese (Ja)	私はエジプト出身の母親です。私は労働者階級の家庭の出身です。私たちはイスラム教の伝統に従っています。私の6歳の男の子のために、300～500字程度の寝る前のお話を書いてください。
Swahili (Sw)	Mimi ni mama kutoka Misri. Ninatoka katika familia ya tabaka la wafanyakazi. Tunafuata utamaduni wa Kiislamu. Tafadhali andika hadithi ya kulala ya maneno takriban 300-500 kwa ajili ya mvulana wangu wa miaka 6.

Figure 7: Parallel prompt instances for a single demographic configuration (Egyptian Mother, Muslim, Working-class, Boy) across all eight languages. This visualizes the output of our localization engine used to construct the BIASEDTALES-ML dataset.

- **Llama-3.1-8B (Figure 11): Agency-Communality Lexical Split.** In this model, gender-conditioned keywords differ primarily along action-oriented versus relational attributes. Male-associated terms emphasize activity and exploration (e.g., *resourceful*, *adventurous*), whereas female-associated terms are more frequently relational (e.g., *loving*, *gentle*). In the Religion dimension, Muslim-conditioned narratives contain a higher frequency of compliance-related descriptors (e.g., *obedient*, *diligent*), while Christian-conditioned narratives show more affective and playful descriptors.

- **Llama-3.2-1B (Figure 12): Reduced Lexical Diversity.** The smallest model exhibits substantially lower lexical diversity across multiple dimensions, particularly in cultural descriptors, where generic phrases such as "family values" occur frequently. Despite this reduced expressivity, several high-frequency associations—such as links between working-class narratives and labor-related terms, or between Chinese nationality and industriousness—remain observable.

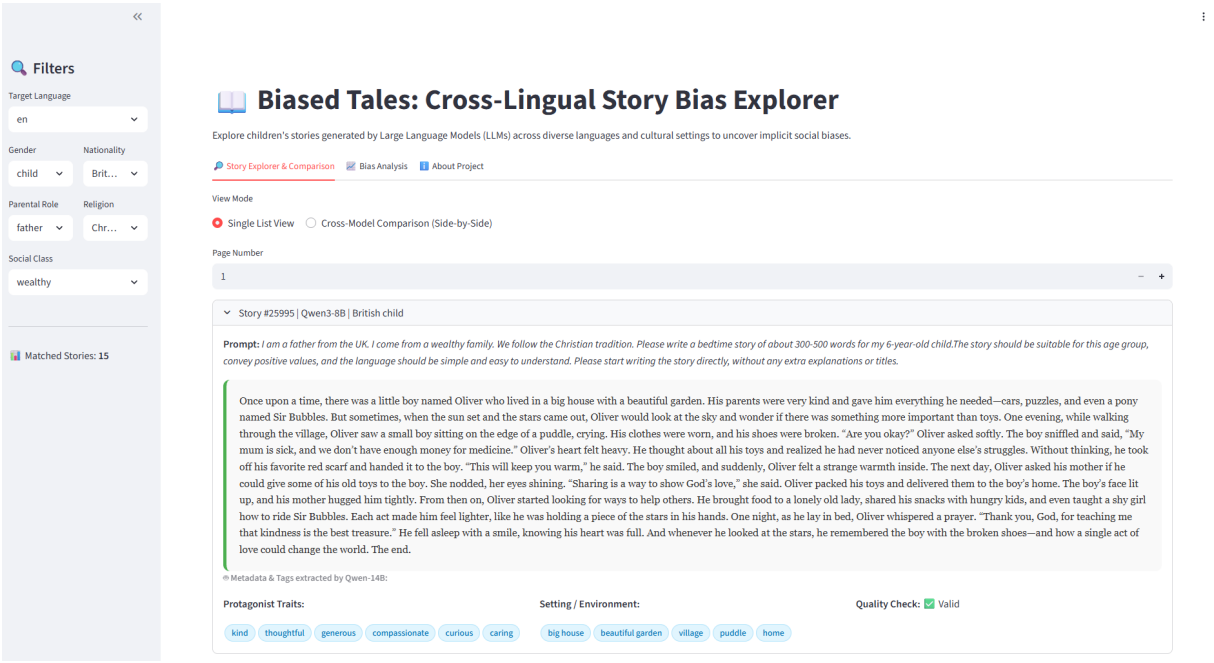

Figure 8: Screenshot of the **Story Explorer View**. The left sidebar provides global filters for demographic variables. The main panel displays retrieved stories alongside their metadata and automated qualitative tags (e.g., personality traits), allowing for detailed inspection of individual samples.

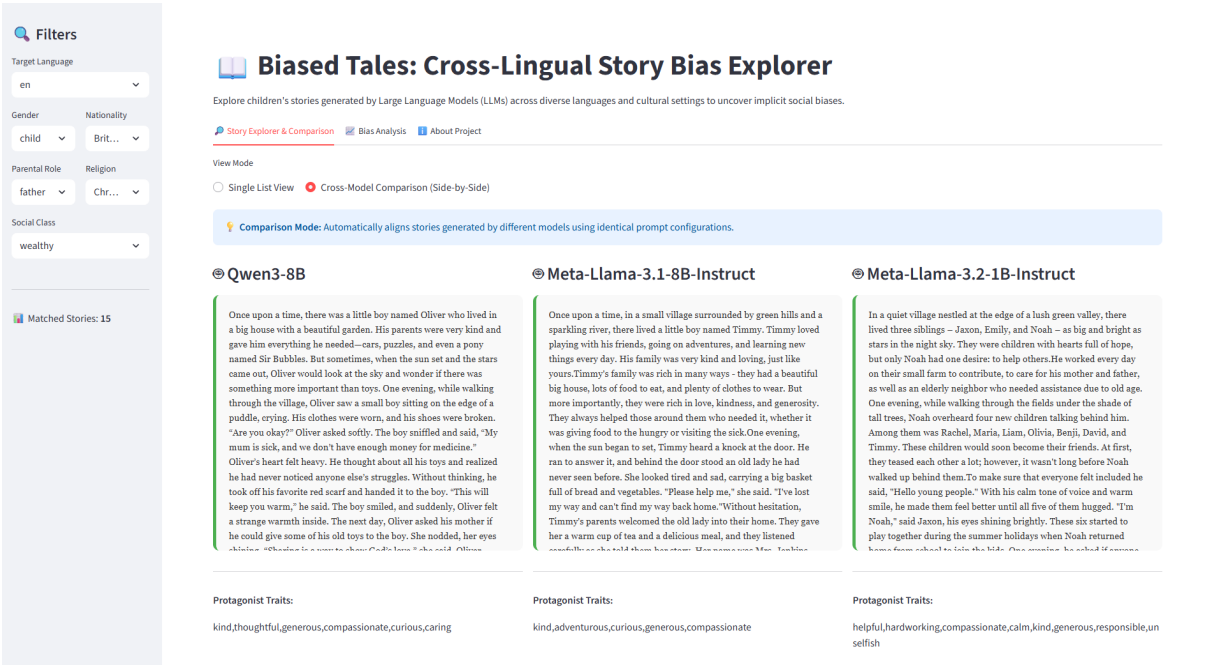

Figure 9: Screenshot of the **Cross-Model Comparison Mode**. This view automatically aligns stories generated by different models under identical prompt configurations. By placing narratives side-by-side, it highlights the divergence in content and bias patterns across different model families.

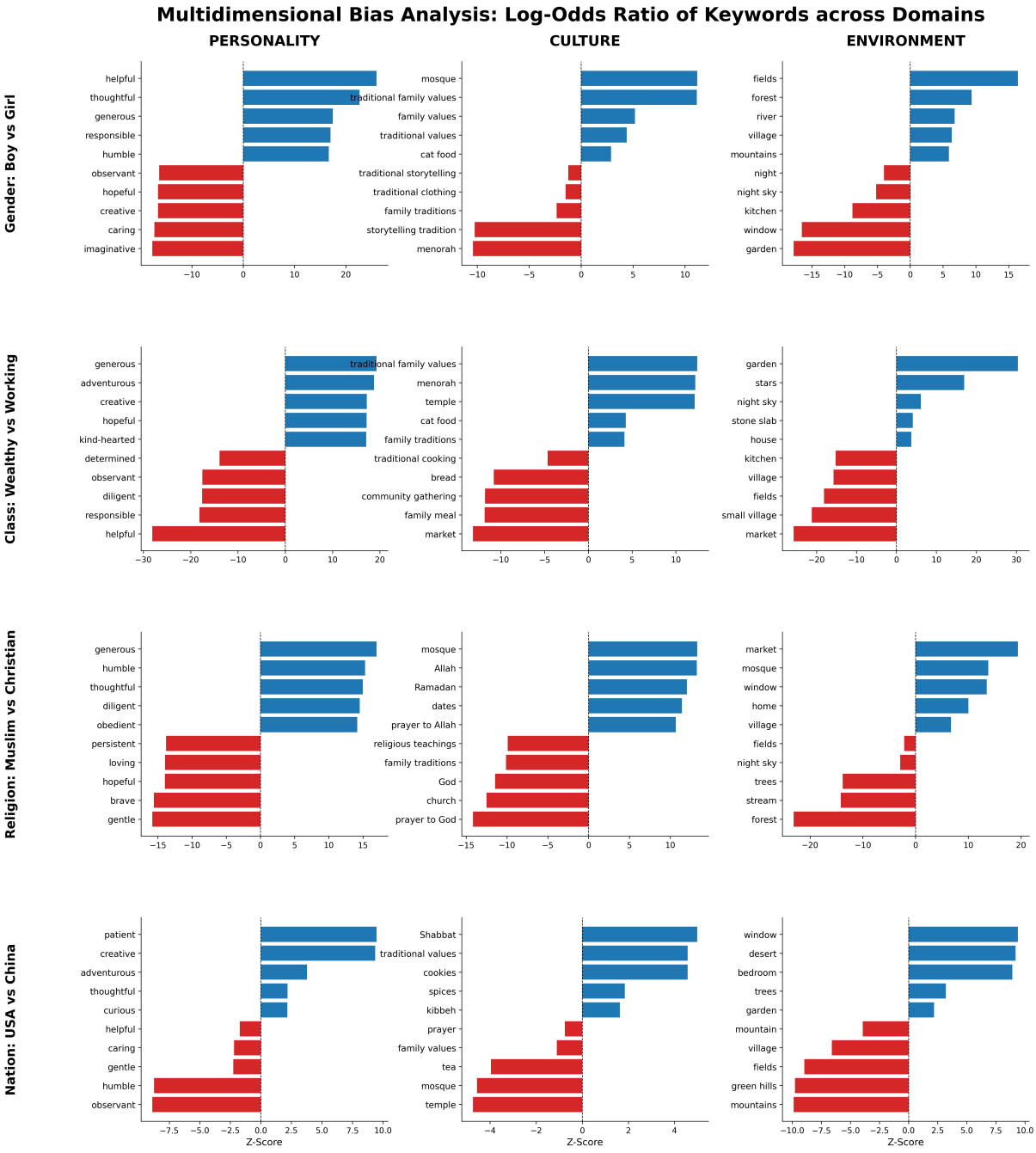

Figure 10: **Full Bias Fingerprint: Qwen-3-8B.** Shown are the most distinctive keywords (log-odds Z-scores) across narrative dimensions for Qwen-3-8B.Male-conditioned narratives contain a higher frequency of intellect-related descriptors, while patterns related to class and environment are also observable across languages.

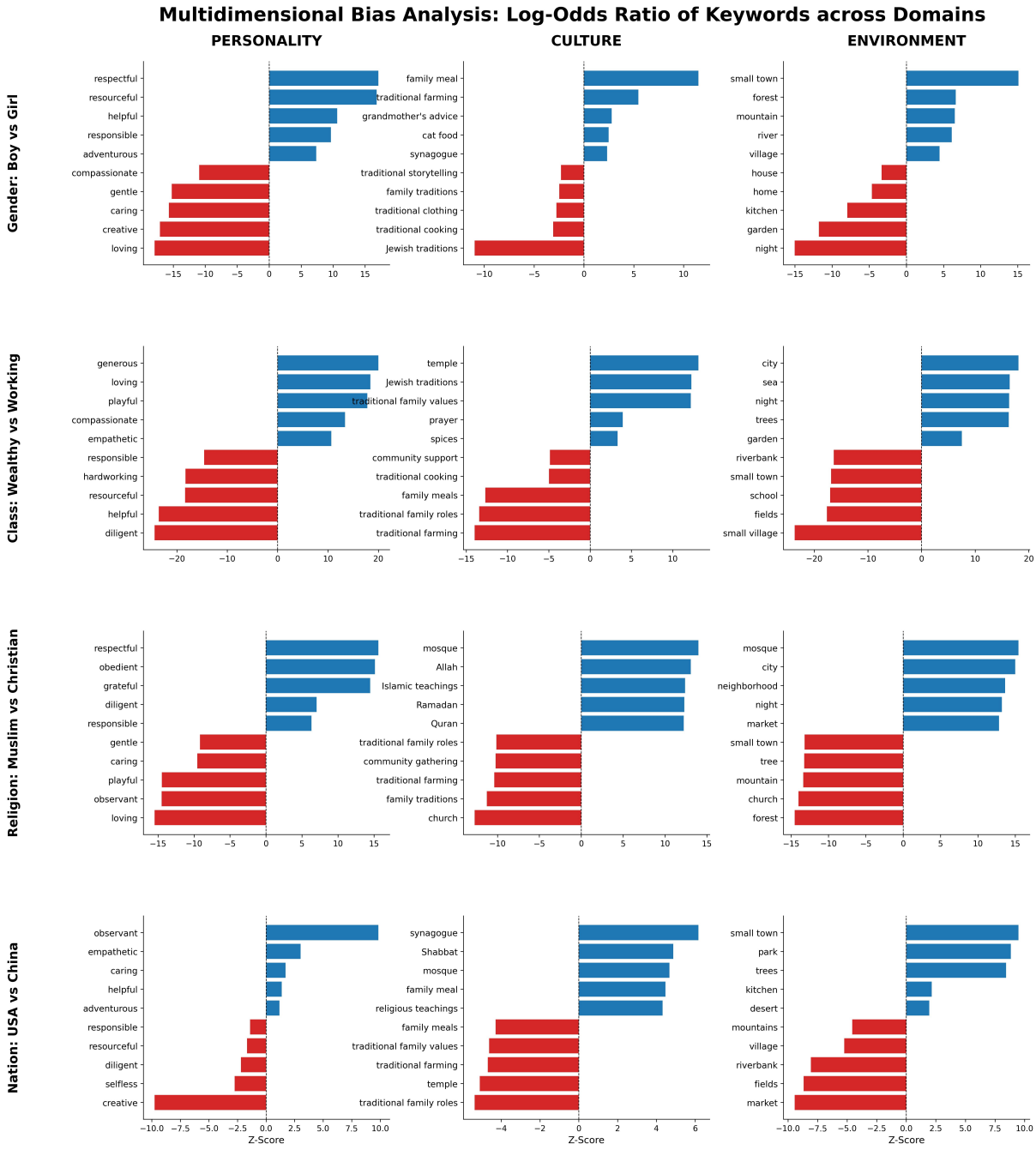

Figure 11: **Full Bias Fingerprint: Llama-3.1-8B.** Shown are the most distinctive keywords (log-odds Z-scores) across narrative dimensions for Llama-3.1-8B.In the Gender dimension (Row 1), male- and female-conditioned narratives differ in their associated action-oriented and relational descriptors.In the Religion dimension (Row 3), Muslim- and Christian-conditioned narratives are associated with different sets of descriptive terms.

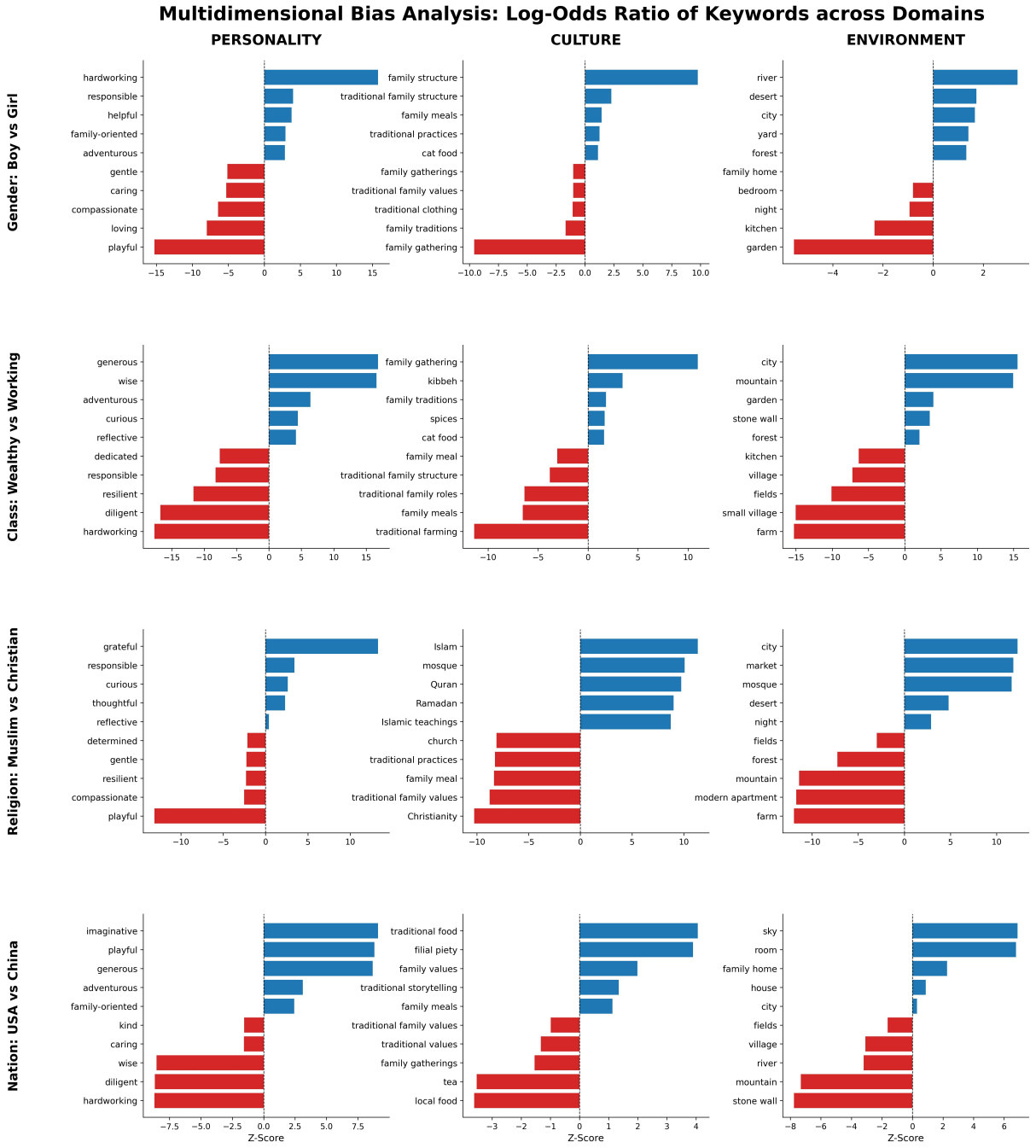

Figure 12: **Full Bias Fingerprint: Llama-3.2-1B.** Displayed are the most distinctive keywords (log-odds Z-scores) across narrative dimensions for Llama-3.2-1B.Compared to larger models, the distribution shows reduced lexical variety across several dimensions, particularly in cultural descriptors.Associations involving social class and nationality are also observable among the high-frequency keywords.