When English Shapes Alterations: Suffix Bias in the Italian Fable Genre

Anonymous ACL submission

Abstract

Despite recent advances in multilingual Large Language Models (LLMs), prior work has identified systematic differences between human-authored and LLM-generated Italian texts, particularly in language-specific morphology. We investigate the "naturalness" of LLM outputs in Italian through the use of alterative suffixes, a productive feature of Italian morphology, analyzing both human-written and LLM-generated fables. Our results show that English-centered models produce fewer suffixes and lower lexical variability than human authors. We further study suffix generation in Machine Translation (MT) using a parallel corpus of English-Italian fables, finding that LLMs rarely reproduce the suffixal strategies used in the Italian translation. We also introduce SUFFIXBIASMT, a contrastive benchmark for evaluating biases against Italian alterative morphology in MT metrics. Experiments reveal that several widely used metrics systematically prefer translations without suffixation, highlighting broader evaluation biases affecting morphologically rich languages like Italian.

1 Introduction

Recent years have seen major efforts to develop multilingual Large Language Models (LLMs) for non-English languages (Orlando et al., 2024; Martins et al., 2024; Ali et al., 2025). However, linguistically oriented studies have identified systematic differences between human-authored and LLM-generated texts, especially in Italian (Cicero, 2023, 2025; Raus, 2025; Antonelli, 2025; Fiorentino and Tavosanis, 2024; Tavosanis, 2024; De Cesare, 2023, 2025, 2026), particularly in the reproduction of language-specific morphological features.

This gap may stem from both data and architectural biases. Multilingual LLMs are typically trained on corpora heavily dominated by English, with other languages are underrepresented (Blevins and Zettlemoyer, 2022; Grattafiori et al., 2024), leading to English-centric behaviors in non-English generation (Guo et al., 2025). Moreover, core architectural components such as tokenization and vocabulary design are often optimized for English, limiting their ability to capture the morphological richness of languages like Italian (Moroni et al., 2025b).

In this work, we focus on the use of Italian alterative suffixes, a productive morphological feature, in Italian fables. Fables, as a genre closely associated with child-centered communication (Dressler and Barbaresi, 1994), provide a particular fertile ground for the use of alterative morphology in both Italian and English. However, the two languages differ substantially in this respect: English possesses a much more limited and less productive system of alterative suffixation than Italian (Dressler and Barbaresi, 1994).

We investigate LLMs' English-centric biases in suffixed word usage in Italian fables through three research questions: i) *How do LLMs handle alterative suffixes in fable generation compared to humans?*, ii) *How does suffix generation emerge when translating fables from English to Italian?*, and iii) *Are MT evaluation metrics biased against suffixed forms in Italian?*

To answer the first question, we compare human-written fables with texts generated by seven LLMs. We find that models pretrained predominantly on English produce significantly fewer alterative suffixes than models more oriented toward Italian. Across models, suffix usage is also concentrated on a narrow set of lexical bases, indicating lower variability and creativity than in human writing.

We then investigate English-Italian literary translation using a parallel corpus of fables translated by a professional translator. Results show that LLMs rarely reproduce the alterative morphology found in human translations, revealing persistent limitations in modeling morphologically expressive forms under translation constraints.

Finally, inspired by recent work on evaluation biases in MT (Perrella et al., 2024; Ahmadi et al., 2025), we introduce SUFFIXBIASMT, a contrastive benchmark designed to assess biases against Italian alterative morphology. We show that several state-of-the-art MT evaluation metrics systematically penalize suffixed forms in favor of their adjective-based counterparts, revealing a source of evaluation bias in Italian and, more broadly, in morphologically rich languages.[1]

2 Related Work

2.1 Manual Evaluation of Italian Synthetically Generated Texts

Linguistic analyses of LLM-generated text in free-form settings have provided useful insights in the Italian research landscape (De Cesare, 2023; Antonelli, 2025). De Cesare (2026) studies gender bias in short biography generation for Italian and French anthroponyms, showing that men are more often referred to by surname, while women are more frequently mentioned by full name or first name alone. Tavosanis (2024) highlights the importance of human evaluation for identifying subtle issues in synthetic text and report actionable findings in Italian-generated corpora. However, little work has properly examined morphological and suffixal phenomena in Italian synthetic child-centered storytelling.

2.2 Baby speech in NLP

Recent work has explored language modeling under child language acquisition constraints (Warstadt et al., 2023; Dhole, 2026). The BabyLM challenge (Warstadt et al., 2023) promotes training on limited, developmentally plausible corpora (10M–100M words), approximating early linguistic input. Follow-up work extends this to multimodal child-centered settings (Hu et al., 2024) and child-directed reasoning benchmarks (Dhole, 2026). Despite this progress, little work analyzes morphological realization in child-directed or storytelling settings for morphologically rich languages such as Italian, where suffixation plays a central role.

2.3 Biases in Machine Translation

Prior work identifies several biases in machine translation. Savoldi et al. (2021) study gender bias in MT, while Xu et al. (2026) show that cultural familiarity of named entities affects translation quality. Recent studies also reveal biases in MT evaluation metrics: Perrella et al. (2024) find that neural metrics like COMET (Rei et al., 2020) are sensitive to surface properties such as sentence complexity, and Ahmadi et al. (2025) show a preference for loanwords over native forms. However, no prior work examines morphological biases in child-centered settings, particularly for Italian suffixation.

3 RQ1: "How do models handle suffix words compared to humans in fable generation?"

3.1 Methodology

To investigate how LLMs produce suffix-based words compared to human-written texts, we first collected 200 Italian fables from human writers (more details in Appendix A). From this human-written fable collection we manually tag all the words that use any of the following eight suffixes: three diminutives, *-ino*, *-etto*, *-ello*; two augmentatives, *-one*, *-asso*; the elative suffix *-issimo*; and two polyfunctional suffixes, *-otto*, *-ozzo*, which can be used as diminutives or augmentatives. We then prompt LLMs to generate Italian fables, and tag them as above.

3.2 Experimental Setup

Models: Following Moroni et al. (2025a), we group instruction-tuned models into three categories: (i) closed models: GPT-3.5-Turbo (Brown et al., 2020) and GPT-4o-mini (OpenAI et al., 2024); (ii) open-data models: Minerva-7B (Orlando et al., 2024), OLMo-2-7B (OLMo et al., 2025), and Teuken-7B (Ali et al., 2025); and (iii) Italian-adapted models: Llama-3.1-8B (Grattafiori et al., 2024), further adapted to Italian through continued pre-training (Llama-LAPT), and its vocabulary-adapted variant with an Italian-optimized tokenizer (Llama-Adapt) (Moroni et al., 2025b).

Generation strategy: We used the `vLLM` library (Kwon et al., 2023) with `temperature` $= 0.7$ for all generations. Synthetic fables were produced by repeating the prompting process 25 times. Prompts used in fable generation are reported in Appendix B.

[1] Code and data are available at `omitted.link`.

Model	Suffixed words	Entry words
Human	93.6	27.7
GPT-4o-mini	39.5	8.2
GPT-3.5-turbo	**157.1**	14.1
OLMO-2-7B	16.8	8.9
Minerva-7B	67.0	**15.8**
Teuken-7B	104.5	12.8
Llama-LAPT	106.1	9.7
Llama-Adapt	107.2	12.4

Table 1: Number of suffix words and entry words in human and synthetic fables, normalized over 10,000 words. Entry words indicate distinct lexical bases appearing with at least one suffix form.

Model	*-asso*	*-ello*	*-etto*	*-ino*	*-issimo*	*-one*	*-otto*	*-ozzo*	# Suff.
Human	0.0	4.3	19.9	**45.7**	26.5	2.8	0.8	0.0	6
GPT-4o-mini	0.0	1.5	10.2	**57.7**	28.5	0.0	2.2	0.0	5
GPT-3.5-turbo	0.0	1.2	44.7	**47.1**	6.6	0.3	0.0	0.0	5
OLMo-2-7B	0.0	6.7	13.3	**46.7**	26.7	6.7	0.0	0.0	5
Minerva-7B	0.0	2.8	**37.5**	26.4	30.6	0.0	2.8	0.0	5
Teuken-7B	0.0	2.5	11.5	**63.1**	22.5	0.4	0.0	0.0	5
Llama-LAPT	0.0	0.6	42.9	**46.6**	9.8	0.0	0.0	0.0	4
Llama-Adapt	0.0	4.4	**53.0**	31.5	9.9	1.1	0.0	0.0	5

Table 2: Row-normalized distribution of Italian suffix types expressed as percentages. **Bold** values indicate row-wise most frequent suffixes.

3.3 Results

Table 1 shows the number of suffix words and entry words (i.e., base forms occurring with at least one suffix) in human and synthetic fables, normalized over 10,000 words. Results show substantial variation across models. Some LLMs generate more suffixed words than humans, while others produce considerably fewer. For example, GPT-3.5-turbo produces the highest number of suffixed words (157.1), whereas GPT-4o generates only 39.5, well below the human average (93.6). Despite this variability, all models exhibit lower lexical diversity than humans. Human texts present 27.7 entry words, while the best-performing model, Minerva-7B, reaches only 15.8, suggesting that models reuse a limited set of lexical bases for suffix formation. Models with stronger Italian exposure during pretraining generally produce more suffixes and greater lexical variety. Teuken-7B and Llama-based models generate suffixed word frequencies close to human texts, whereas the English-centric OLMo-2-7B produces the fewest suffixed and entry words, highlighting the impact of pretraining data composition on Italian suffixation. Table 2 shows the distribution of suffix types. Human fables are dominated by *-etto*, *-ino*, and *-issimo*, a pattern partially reproduced by most models, indicating that LLMs capture some human-like preferences in suffix selection.

4 RQ2: "How does suffix generation emerge when translating fables from English to Italian?"

4.1 Methodology and Setup

To further analyze how suffixed words are handled in generative contexts we investigate how LLMs produce them when translating English fables into Italian. We considered 68 fables from *"Favole al Telefono"* by Gianni Rodari, which have been translated into English by a professional translator. We then prompted models, first, with an Italian system prompt and, then, with an English system prompt to determine the impact of crosslingual prompting in an MT setting. We followed the experimental setup defined in Section 3, further details about MT prompt strategies are reported in Appendix C.

4.2 Results

Table 3 reports English-to-Italian machine translation performance using both Italian and English prompts. Translation quality is measured with COMET, while suffixed accuracy (*Suf. Acc.* in the Table) denotes the percentage of suffixed words from the gold Italian translation correctly reproduced by the model. Prompting language affects performance, although no consistent trend emerges across models. GPT-4o-mini and GPT-3.5-turbo achieve the highest COMET scores (around 85), whereas OLMo-2-7B and Minerva-7B perform substantially worse. These latter models also obtain the lowest suffix accuracy, suggesting a relation between translation quality and the ability to reproduce Italian suffixation. Overall, all models struggle to generate gold suffixed forms. The best suffix accuracy is achieved by GPT-3.5-turbo with English prompting (18.5) and GPT-4o-mini with Italian prompting (16.9), indicating that suffix reproduction remains difficult even for strong translation models.

5 RQ3: "Are MT evaluation metrics biased against suffixed forms in Italian?"

5.1 Methodology

MT evaluation metrics are known to exhibit systematic biases toward lexical and stylistic variation

Model	Italian		English	
	Comet	Suf. Acc.	Comet	Suf. Acc.
gpt-4o-mini	**85.1**	**16.9**	**85.0**	13.2
gpt3.5-turbo	82.4	11.3	84.6	**18.5**
OLMo-2-7B	66.9	7.0	68.0	5.4
Minerva-7B	76.8	5.8	75.5	6.6
Teuken-7B	82.7	12.5	81.2	10.1
Llama-3.1-8B-LAPT	84.1	9.8	84.0	13.2
Llama-3.1-8B-SAVA	82.9	10.1	83.4	11.1

Table 3: Machine translation results from English to Italian using Italian and English system prompts. COMET and *Suf. Acc* are reported, where the latter reports the percentage of gold Italian suffixed words correctly reproduced in the generated translations.

Original Italian	Rewritten Italian	English Translation
La *nuvoletta* dormiva vicino al camino.	La *piccola nuvola* dormiva vicino al camino.	The *little cloud* was sleeping near the fireplace.

Table 4: Example of the controlled rewriting process used to analyze MT metric biases toward Italian suffixed forms.

(Perrella et al., 2024; Ahmadi et al., 2025). To assess their behavior on Italian alterative morphology, we introduce SUFFIXBIASMT, a contrastive benchmark of translation triples.

To construct the benchmark, we extracted sentences from Italian fables written by humans that contain suffixed words (e.g., *nuvoletta*). We then replaced each suffixed form with a semantically equivalent adjective-based paraphrase (e.g., *piccola nuvola*), thereby creating an alternative Italian version of the sentence. Finally, we translated the adjective-based variants into English. Table 4 illustrates the transformation process.

SUFFIXBIASMT contains 453 items. Thus, each instance in SUFFIXBIASMT consists of: (i) an English sentence $s_{\textbf{eng}}$, (ii) an Italian translation using alterative suffixation $t_{\textbf{suff}}$, and (iii) a semantically equivalent translation without suffixation $t_{\textbf{no-suff}}$.

5.2 Experimental Setup

We evaluate reference-free metrics, including COMET (Rei et al., 2020), xCOMET (Guerreiro et al., 2023), and COMETKiwi (Rei et al., 2022). Scores are computed using the English source sentence (s_{eng}) paired with either a suffix-preserving Italian translation (t_{suff}) or its non-suffixed variant ($t_{\text{no-suff}}$). We also evaluate sentinel-src (Perrella et al., 2024), a source-only metric predicting translation difficulty, where higher scores indicate easier sentences. Here, Italian is the source language, and we compare t_{suff} and $t_{\text{no-suff}}$ as alternative inputs.

Model	No Suff.	Suff.	Δ
XCOMET-XL	0.793	0.781	0.012
wmt23-cometkiwi-da-xl	0.716	0.698	0.018
wmt22-cometkiwi-da	0.832	0.819	0.013
wmt20-comet-qe-da	0.171	0.146	0.025
sentinel-src	0.102	0.068	0.034

Table 5: Reference-free MT evaluation scores for Italian translations with and without suffixed forms. Δ represents the score difference between non-suffixed and suffixed variants, where positive values indicate a preference for non-suffix constructions.

5.3 Results

Table 5 reports the scores and the performance gap (Δ) between suffixed and non-suffixed variants. All metrics consistently assign higher scores to translations without suffixes, favoring syntactic paraphrases with adjectives over morphologically altered forms. These results suggest that current MT evaluation metrics exhibit biases against Italian suffixation, reflecting broader English-centric preferences in neural evaluation systems.

6 Conclusions

We presented the first study on English-centric biases in the generation and translation of Italian alterative morphology in fables. Across both free-form generation and English-Italian translation, LLMs consistently underproduce alterative suffixes compared to human authors, favoring a narrower and less varied lexical repertoire. These findings suggest that current multilingual LLMs still struggle to model morphologically expressive phenomena that are highly productive in Italian but marginal in English. To further investigate this issue, we introduced SUFFIXBIASMT, a contrastive benchmark targeting evaluation biases toward Italian alterative morphology. Our results show that several widely used MT evaluation metrics systematically penalize suffixed forms in favor of adjective-based alternatives, highlighting an overlooked source of bias in automatic evaluation for morphologically rich languages. Overall, our work emphasizes the importance of moving beyond surface fluency when evaluating multilingual LLMs and MT systems, and calls for more linguistically informed modeling and evaluation practices for non-English languages.

Limitations

Despite the comprehensiveness of the proposed work, several limitations should be acknowledged.

First, the analysis is restricted to a specific genre of child-related fables. While this provides a controlled setting for investigation, future work should broaden the scope to include additional genres and stylistic registers in Italian literature, in order to better capture potential variation in morphological patterns.

Second, for reasons of simplicity and the substantial effort required for annotation, this study focuses exclusively on Italian. This language-specific scope limits the generalizability of the findings. Extending the analysis to other languages would be valuable, particularly those in which English-centric morphological biases may emerge in downstream applications.

Finally, due to computational constraints, the experiments are limited to open-weight, open-data models with up to 8B parameters. While this enables reproducibility and accessibility, it restricts the evaluation of larger-scale models. Future research could investigate whether the observed effects persist or change when scaling to larger architectures.

References

Sina Ahmadi, Micha David Hess, Elena Álvarez-Mellado, Alessia Battisti, Cui Ding, Anne Göhring, Yingqiang Gao, Zifan Jiang, Andrianos Michail, Peshmerge Morad, Joel Niklaus, Maria Christina Panagiotopoulou, Stefano Perrella, Juri Opitz, Anastassia Shaitarova, and Rico Sennrich. 2025. ConLoan: A contrastive multilingual dataset for evaluating loanwords. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30070–30090, Vienna, Austria. Association for Computational Linguistics.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Jan Ebert, Alexander Arno Weber, Richard Rutmann, Charvi Jain, Max Lübbering, Daniel Steinigen, Johannes Leveling, Katrin Klug, Jasper Schulze Buschhoff, Lena Jurkschat, Hammam Abdelwahab, Benny Jörg Stein, Karl-Heinz Sylla, Pavel Denisov, Nicolo' Brandizzi, Qasid Saleem, and 22 others. 2025. Teuken-7b-base & teuken-7b-instruct: Towards european llms. *Preprint*, arXiv:2410.03730.

Giuseppe Antonelli. 2025. Storia brevissima (ma molto intensa) dell'ia-taliano. *Lingue e culture dei media*, 9(1):4–50.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Francesco Cicero. 2023. L'italiano delle intelligenze artificiali generative. *Italiano LinguaDue*, 15(2):733–761.

Francesco Cicero. 2025. Esercizi di stile. la narrativa per l'infanzia delle intelligenze artificiali. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses*, 2(2).

Anna-Maria De Cesare. 2023. Assessing the quality of chatgpt's generated output in light of human-written texts: A corpus study based on textual parameters. *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos*, 10:179–210.

Anna-Maria De Cesare. 2025. Llm referential chain generation.: A qualitative case study based on italian biographies produced by gpt-4. *Linguistik Online*, 136(4):25–52.

Anna-Maria De Cesare. 2026. Gender biases in gpt-4 short biographies. : A corpus study on italian and french anthroponyms. *Linguistik Online*, 144(3):227–245.

Kaustubh D. Dhole. 2026. Babyreasoningbench: Generating developmentally-inspired reasoning tasks for evaluating baby language models. *Preprint*, arXiv:2601.18933.

Wolfgang U. Dressler and Lavinia M. Barbaresi. 1994. *Morphopragmatics*. De Gruyter Mouton, Berlin, New York.

G. Fiorentino and M. Tavosanis. 2024. Chiaro, sintetico, e brillante: l'italiano dei testi redatti con l'ia funziona? *LId'O*, 5:37–65.

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 4 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Preprint*, arXiv:2310.10482.

Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2025. Do large language models have an English accent? evaluating and improving the naturalness of multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3823–3838, Vienna, Austria. Association for Computational Linguistics.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Luca Moroni, Javier Aula-Blasco, Simone Conia, Irene Baucells, Naiara Perez, Silvia Paniagua Suárez, Anna Sallés, Malte Ostendorff, Júlia Falcão, Guijin Son, Aitor Gonzalez-Agirre, Roberto Navigli, and Marta Villegas. 2025a. Multi-LMentry: Can multilingual LLMs solve elementary tasks across languages? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34126–34157, Suzhou, China. Association for Computational Linguistics.

Luca Moroni, Giovanni Puccetti, Pere-Lluís Huguet Cabot, Andrei Stefan Bejgu, Alessio Miaschi, Edoardo Barba, Felice Dell'Orletta, Andrea Esuli, and Roberto Navigli. 2025b. Optimizing LLMs for Italian: Reducing token fertility and enhancing efficiency through vocabulary adaptation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6661–6675, Albuquerque, New Mexico. Association for Computational Linguistics.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 24 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and Irwan Bello et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy. CEUR Workshop Proceedings.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244, Bangkok, Thailand. Association for Computational Linguistics.

Rachele Raus. 2025. *Inclusione ed elaborazione del linguaggio naturale nell'era dell'intelligenza artificiale generativa*. Zenodo.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. *Preprint*, arXiv:2209.06243.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Mirko Tavosanis. 2024. Valutare la qualità dei testi generati in lingua italiana. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses*, 1(1).

Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023. Call for papers – the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *Preprint*, arXiv:2301.11796.

Lu Xu, Luca Moroni, and Roberto Navigli. 2026. Cultural and knowledge biases in llms through the lens of entity-aware machine translation. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 8794–8812, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

A Human Fables Collection

Table 6 reports the human-authored fable collections used in our analysis. We organize the corpus into two subsets according to the target reader age: (i) *3–5 years*, comprising fables intended to be read aloud by adults but primarily designed for early childhood audiences; and (ii) *6–10 years*, comprising texts intended for independent reading by young readers.

The categorization was performed manually by a trained annotator with a background in linguistics, based on discourse, syntactic, and morphological cues.

We relied on the copyrighted fable texts solely for release purposes. We do not share complete fables or full paragraphs from them.

B Fables Generation Prompts

In Table 7 we report the prompt used to generate free-form children's fables. To increase variability, we condition the prompt on the target reader's age, distinguishing between (i) 3–5 years and (ii) 6–10 years, following the human-written resources described in Appendix A.

C Machine Translation Prompts

In Tables 8 and 9, we report the Italian and English prompts used to evaluate models on fable translation (from English into Italian).

Human written fables	N. of fables	Publication date
3–5 years		
Le Favole di Morfeus	19	2018
Letture per bambini	18	2016
Storie brevi della buonanotte	37	2015–2022
Favole per bambini	6	2020
15 brevi racconti per bambini sulla natura	10	2021
Falcone, Favole della buonanotte	10	2016
6–10 years		
Rodari, Favole al telefono	69	1995 [1962]
Falchi, Fiabe illustrate per bambini	10	2011
Iannicello, Racconti e favole per bambini	4	2017
Coletta (a cura di), C'erano una volta... le Favole	14	2008
Romani, Favole per bambini	3	2020–2021
Tot.	**200**	

Table 6: Composition of the HUM fable subcorpus grouped by age range.

SYSTEM
`Il tuo compito è generare testi narrativi.`
USER
`Scrivi una favola per bambini {x}`

Table 7: Prompt for fables generation. The parameter `span_range` is defined as `(3-5 anni)` and `(6-10 anni)`.

SYSTEM
`Il tuo compito è tradurre storie dall'Inglese all'Italiano. Mantieni tono, significato e stile originali. Rispondi solo con la storia tradotta in Italiano.`
USER
`Storia: {fable}`

Table 8: Prompt for fables translation in Italian.

SYSTEM
`Your task is to translate stories into Italian. Preserve the original tone, meaning, and style. Reply only with the translated story in Italian.`
USER
`Story: {fable}`

Table 9: Prompt for fables translation in English.