Evaluating LLM Style Transfer Through Readability-Based Age Assessments

Maria Di Maro[1], **Antonio Origlia**[1], **Leonilda Bilo**[2],
Roberta Meo[3], **Pietro Maturi**[4], **Francesca Nappo**[2]

[1]Dept. of Electrical Engineering and Information Technology - University of Napoli Federico II,
[2]Dept. of Neuroscience, Reproductive Sciences and Dentistry - University of Napoli Federico II,
[3] Neurology Outpatient Service, Napoli 1 Health District,
[4]Dept. of Social Sciences - University of Napoli Federico II

Abstract

Adaptability to the audience is an important feature for conversational systems, especially in the healthcare dissemination field, where scientific concepts have to be delivered to a potentially wide range of users. This work presents an evaluation of the capability of LLMs to support style transfer according to the target user's age group. Two complementary evaluation methods were employed: an automatic assessment based on the ARI readability index, and a human experts evaluation focusing on appropriateness depending on the user's educational level as well as content accuracy. Results show that LLMs efficiently switch style when provided with information about the user's age while managing content still requires the adoption of safety measures.

1 Introduction

One important application of Conversational AI is public health dissemination. The World Health Organization (WHO) defines health literacy as 'the ability to access, comprehend, evaluate and communicate information to promote, maintain and improve health in a variety of settings across the life course" (Rootman and Gordon-El-Bihbety, 2008). Low levels of health literacy are associated with public health risks and inequalities in access to medical services (Ratzan and Parker, 2000), reduced engagement with healthcare (Kobayashi et al., 2015), and difficulties in understanding medical instructions (Miller, 2016).

Large Language Models (LLMs) have become increasingly accessible and are frequently adopted in Conversational AI systems, including applications related to health information dissemination. Despite their strong generative capabilities, LLMs are known to suffer from hallucinations and semantic inaccuracies, which pose serious challenges in high-risk domains such as healthcare. Consequently, their use in medical contexts requires careful and task-specific evaluation, rather than an assumption of reliability.

A field in which their generative capabilities find broad expression is text style transfer. This task refers to the possibility to transform an input text into a desired style without compromising the semantics. Such capability is important when considering conversational AI systems and their possibility to adapt to the audience by adopting a specific language style. In Pu and Demberg (2023), this ability to shift sentences between formal and informal registers was assessed (Rao and Tetreault, 2018). While results show that the model can follow stylistic instructions, its outputs differ systematically from human-written texts, highlighting the importance of evaluating stylistic and linguistic properties even when semantics appear preserved.

A rigorous evaluation of LLMs serves multiple purposes, including identifying strengths and limitations, informing effective human–LLM interaction strategies, and supporting the analysis of emergent abilities (Chang et al., 2024). While LLMs perform well in several natural language understanding tasks, such as sentiment analysis (Liang et al., 2022; Zeng et al., 2022) and text classification (Yang and Menczer, 2025), they show limitations in natural language inference (Lee et al., 2023) and semantic understanding (Riccardi et al., 2024). In natural language generation, they achieve strong results in summarization and translation (Bang et al., 2023), although performance is generally higher when translating into English than into other languages (Bang et al., 2023; Lyu et al., 2023). In dialogue, despite their apparent fluency, LLMs may underperform compared to systems specifically designed for particular tasks (Bang et al., 2023).

In this paper, we present an exploratory study on the adaptability of an LLM for style transfer of scientific health-related texts with respect to the age of the intended audience. We specifically focus on adaptations aimed at children and young adults,

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 307–311
February 26–March 1, 2026. ©2026 Association for Computational Linguistics

as possible non-expert targets of such dissemination, assessing whether such transformations improve age-appropriate readability while preserving content integrity. We conducted the evalutation using only one model to enable an in-depth, exploratory evaluation of age-oriented style transfer in healthcare, rather than a broad model comparison, combining automated readability-based age assessments with human expert evaluations. This work is therefore intended as a first step toward more extensive analyses involving multiple models and domains, which we plan to address in future work. The following sections describe the corpus, methodology, and results in detail.

2 Methods and materials

To linguistically evaluate the outputs generated by LLMs, we relied on well-established readability metrics from the literature. The Supporting Patients with Embodied Conversational Interfaces and Argumentative Language (SPECIAL) project[1], which aims to identify and counteract stigma and prejudice about epilepsy through conversational AI, is based on a knowledge base of scientific and informative texts provided by authoritative Italian and international sources, such as Lega Italiana Contro l'Epilessia (LICE) and International League Against Epilepsy (ILAE). These were analysed in terms of linguistic features, including syntactic complexity, use of specialised terminology, and readability, to understand the differences between the two text types, in line with previous studies (Sabatini, 1999).

Specifically, readability tests, which will be described section 2.1, confirmed that scientific texts, in both Italian and English, are generally more complex than informative texts. However, important differences emerged between the two languages: English scientific texts exhibited higher readability than Italian oneswhile English informative texts achieved high readability scores, designed to reach broader audiences and facilitate access to complex information. Finally, although scientific texts exhibit more specialised language, they partially overlap with informative texts; thus, it is more appropriate to view the two genres along a continuum shaped by factors such as publication venue, author expertise, and intended audience. With regard to the last point, we decided to investigate this matter by focusing our analysis on the application of one specific readability test which considers the age of the audience. This was then compared with a human evaluation.

Concerning the corpus, 21 Italian scientific texts were selected and processed using LLaMA 3.3, which was tasked with adapting each text to a specific age group based on the readability index (8–9, 9–10, 10–11, 11–12, 12–13, 13–14, 14–15, 15–16, 16–17, 17–18, 18–22). The model received the prompt: "Based on this text, write an original popular science text in Italian conveying the information from the source. The text should be understandable by individuals aged [target age group]. Do not summarize; generate an original text." In total, 21 texts were adapted across 11 age groups, resulting in a corpus of 231 texts.

2.1 Readability Indexes

One of the first work on readability scores was presented by Flesch (1948). The author described a revised system for assessing the comprehension difficulty of written texts using two new formulas considering the following factors: i) average sentence length in words, ii) average word length in syllables, iii) average percentage of personal words (i.e., nouns with gender, pronouns with gender, etc.), and iv) average percentage of personal sentences (i.e., quotations, exclamations, etc.). From this, other scores were formulated. The Flesch-Kincaid Grade Level translates similar measurements into a U.S. school grade level, indicating the minimum education required to understand the text (Thomas et al., 1975). The Gunning Fog Index estimates the years of formal education needed to comprehend a passage, giving additional weight to longer, complex words (Gunning, 1952). The Coleman-Liau Index uses characters per word and sentence length to approximate the U.S. grade level, relying on orthographic rather than syllabic complexity (Coleman and Liau, 1975). Finally, the SMOG Index focuses on the frequency of polysyllabic words to predict the years of education needed for full comprehension (Mc Laughlin, 1969). Most of these, work very well on English but might fail with other language, such as Italian. In Dell'Orletta et al. (2011), a different index is presented which also considers language-dependent aspects (i.e., frequency lexicon). This index combines traditional raw text features with lexical, morpho-syntactic and syntactic information to better capture nuances.

Traditionally, such indexes have been frequently used for evaluating scientific texts, mostly medi-

[1] https://www.specialprojectunina.com/

Score	Age	Grade level
1	5-6	Kindergarten
2	6-7	1st grade
3	7-8	2nd grade
4	8-9	3rd grade
5	9-10	4th grade
6	10-11	5th grade
7	11-12	6th grade
8	12-13	7th grade
9	13-14	8th grade
10	14-15	9th grade
11	15-16	10th grade
12	16-17	11th grade
13	17-18	12th grade
14	18-22	College

Table 1: ARI scores with corresponding age groups and grade levels

cal ones. Recently, they have been also applied to LLMs readability assessments in healthcare, as in Gencer (2024). This study found that ChatGPT's responses on lung cancer are challenging to read, typically at a college level or higher. This poses a concern, as users of varying ages and educational backgrounds may struggle to understand the information, increasing the risk of misinterpretation.

On these premises, we selected one readability index in particular, which considers the age of the reader: the Automated Readability Index (ARI). It estimates the years of education required to understand the text on the first reading. In order to do that, it considers the mean number of characters per word and the mean number of words per sentence within a given text sample. Table 1 reports the age groups corresponding to each score.

2.2 Online questionnaire

While automatic evaluation is time saving, less subjective and more standardised, human evaluation is still more reliable especially in domains where the expert's opinion is fundamental, like in healthcare. Unlikely from automatic evaluation, human evaluation is closer to the actual application scenario and can provide more comprehensive and accurate feedback (Chang et al., 2024). In our work, we used the human evaluation as a further comparable validation for the results collected from the application of ARI to the text analysis. For this experiment, a questionnaire designed in Qualtrics[2] was administered via the LICE community of epileptologists.

From the generated corpus described in the previous section, one Italian text was selected for each age group, along with one scientific text. Participants received the following instructions: *We invite you to take part in a brief experiment lasting approximately 10 minutes. During the activity, you will be asked to read and evaluate four informative medical texts. The questions will focus primarily on the accuracy of the content and on the age group for which the texts appear to be intended.*

Each participant was presented with four texts, each corresponding to a macro-age group: lower (8–11), lower-middle (11–14), upper-middle (14–17), and upper (17–specialist). Texts were randomized to ensure balanced presentation across the experiment. Fifteen expert epileptologists participated in the study and evaluated the texts by assigning a score from 1 to 5 to the following questions: 1) Do you think the text contains accurate information? (from "absolutely not" to "absolutely yes"; participants were also asked to justify low scores) 2) What is the minimum educational level required to understand this text? (primary, lower secondary, upper secondary, university, postgraduate).

3 Results

Automatic evaluation We applied the ARI scores to the texts, using the following formula:

$$\text{ARI} = 4.71 \times \left(\frac{\text{characters}}{\text{words}}\right) + 0.5 \times \left(\frac{\text{words}}{\text{sentences}}\right) - 21.43$$

A Kruskal–Wallis rank sum test was performed to examine whether the ARI scores differed significantly across the text categories (defined by age groups and the "Scientific" category). The test yielded a highly significant result ($p < 2.2e - 16$), indicating that at least some categories differed in their ARI distributions. To identify where these differences occurred, pairwise Wilcoxon rank-sum tests with Bonferroni correction were conducted. The results revealed multiple significant pairwise differences, particularly between the scientific texts and many of the age-based groups, as well as between the lowest age groups (e.g., 8–9, 9–10) and higher-level categories. These results were then grouped into 4 different groups (Section 2.2), low, middle-low, middle-high, and high. As shown in the boxplot in Figure 1, we observe an increasingly complex text style.

Human evaluation The data collected with the human-evaluation questionnaire are divided into two sets of results: those related to style (i.e., the education level appropriate for understanding the presented text) and the qualitative results regarding the correctness of content produced during the style-adaptation phase. As noted, LLMs can be unstable in this respect: changing the style may alter the generated content. Because this can have potentially disastrous consequences in the medical domain, it is important to verify their reliability.

[2] https://www.qualtrics.com/

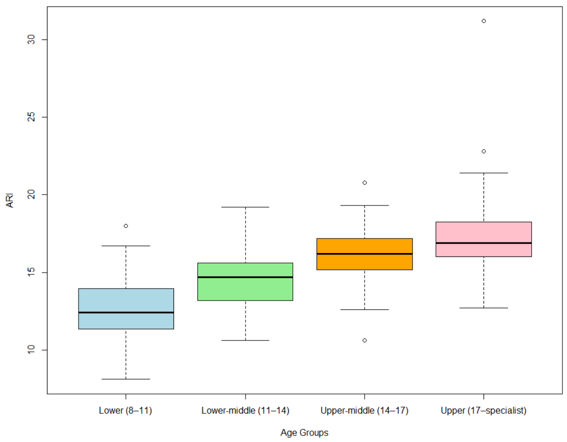

Figure 1: ARI scores across different age groups

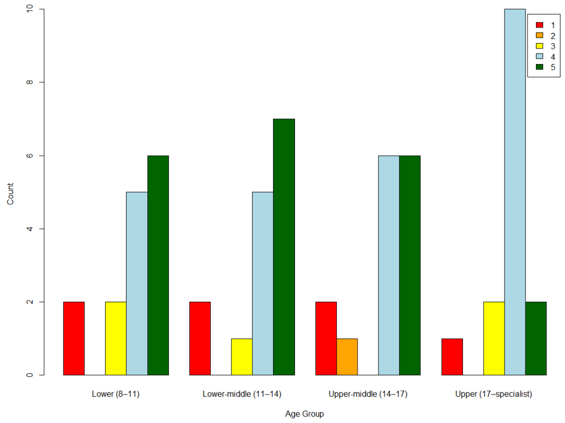

Figure 3: Distribution of Likert scores by Age Group on contents correctness

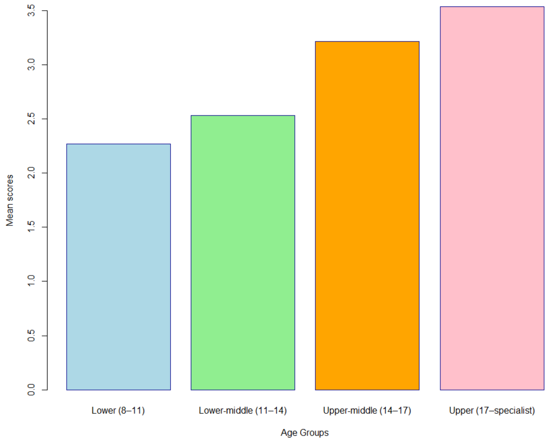

Figure 2: Likert scores concerning the educational level

Regarding the style, the results confirm the automatic analysis: they effectively map a continuum of age groups onto educational levels (primary, lower-secondary, upper-secondary, postgraduate on the y-axis). These findings are shown in Figure 2.

With respect to content correctness, it is noteworthy that mean scores generally do not fall below 3 (age group 16–17) and reach a maximum of 4.6 (age group 14–15). However, correctness tends to decline in average in the older groups, as shown in the barplot in Figure 3. This is understandable: texts for the younger groups are less specialist and therefore less exposed to precision errors, whereas texts for older groups - and especially the scientific texts (which are not LLM-generated) - may contain interpretations that a specialist, disagreeing with a particular reading or coming from a different school, could find imprecise or embreaceables. Nevertheless, as shown in 3, a considerably larger proportion of the scores falls within the light blue bar (score 4), whereas the lower scores are far less represented. A commonly reported issue among the participants was the perception that texts were machine-translated. This complaint applies not only to generated texts but also to the scientific ones. This may reflect the predominance of English in scientific literature, which can make the Italian version read as unnatural or calqued. Among other comments, terminology was sometimes considered incorrect (e.g., “attacks” instead of “seizures”), and several factual inaccuracies were highlighted, such as misleading claims about breastfeeding and physical activity wrt epilepsy. Important clinical guidance was missing or misstated: seizure management instructions were incomplete, diazepam’s role was described inaccurately, and inappropriate advice was given. Legal aspects were also misrepresented, with commentators clarifying that driving restrictions stem from licensing laws.

4 Conclusions

This work shows that LLMs are indeed capable of adapting texts to the needs of users belonging to a specific age group. However, human evaluation also revealed the presence of content-related issues, especially in texts for older age groups, where the level of detail is higher. Although correctness is generally judged as medium to high, in the health domain, even small errors can be disastrous. For this reason, our future works on developing conversational applications to improve the general public’s health literacy will be based on putting stronger control over content generation while leaving form adaptation to LLMs.

5 Acknowledgments

This work is supported by the Supporting Patients with Embodied Conversational Interfaces and Argumentative Language (SPECIAL) project, funded by the University of Naples on the "Fondi per la Ricerca di Ateneo" (FRA) program (CUP: E65F22000050001).

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read–it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):p221 – 233.

Adem Gencer. 2024. Readability analysis of chatgpt's responses on lung cancer. *Scientific Reports*, 14(1):17234.

Robert Gunning. 1952. The technique of clear writing. *(No Title)*.

Lindsay C Kobayashi, Jane Wardle, and Christian von Wagner. 2015. Internet use, social engagement and health literacy decline during ageing in a longitudinal cohort of older english adults. *J Epidemiol Community Health*, 69(3):278–283.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models infer and disagree like humans? *CoRR*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with chatgpt. *arXiv preprint arXiv:2305.01181*.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Tricia A Miller. 2016. Health literacy and adherence to medical treatment in chronic and acute illness: a meta-analysis. *Patient education and counseling*, 99(7):1079–1086.

Dongqi Pu and Vera Demberg. 2023. Chatgpt vs human-authored text: Insights into controllable text summarization and sentence style transfer. *arXiv preprint arXiv:2306.07799*.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

SC Ratzan and RM Parker. 2000. Health literacy. *National library of medicine current bibliographies in medicine. Bethesda: National Institutes of Health, US Department of Health and Human Services*.

Nicholas Riccardi, Xuan Yang, and Rutvik H Desai. 2024. The two word test as a semantic benchmark for large language models. *Scientific Reports*, 14(1):21593.

Irving Rootman and Deborah Gordon-El-Bihbety. 2008. A vision for a health literate canada. *Ottawa, ON: Canadian Public Health Association*, page 50.

Francesco Sabatini. 1999. 'rigidità-esplicitezza'vs 'elasticità-implicitezza': possibili parametri massimi per una tipologia dei testi. *Linguistica testuale comparativa*, pages 142–172.

Georgelle Thomas, R Derald Hartley, and J Peter Kincaid. 1975. Test-retest and inter-analyst reliability of the automated readability index, flesch reading ease score, and the fog count. *Journal of Reading Behavior*, 7(2):149–154.

Kai-Cheng Yang and Filippo Menczer. 2025. Accuracy and political bias of news source credibility ratings by large language models. In *Proceedings of the 17th ACM Web Science Conference 2025*, pages 127–137.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.