ChatGPT in the Classroom

A Preliminary Exploration on the Feasibility of Adapting ChatGPT to Support Children's Information Discovery

Murgia, Emiliana; Abbasiantaeb, Zahra; Aliannejadi, Mohammad; Huibers, Theo; Landoni, Monica; Pera, Maria Soledad

DOI
10.1145/3563359.3597399

Publication date
2023

Document Version
Final published version

Published in
UMAP 2023 - Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization

ChatGPT in the Classroom: A Preliminary Exploration on the Feasibility of Adapting ChatGPT to Support Children's Information Discovery

Emiliana Murgia*
emiliana.murgia@unige.it
Università di Genova
Italy

Zahra Abbasiantaeb
z.abbasiantaeb@uva.nl
University of Amsterdam
The Netherlands

Mohammad Aliannejadi
m.aliannejadi@uva.nl
University of Amsterdam
The Netherlands

Theo Huibers
t.w.c.huibers@utwente.nl
University of Twente
The Netherlands

Monica Landoni
monica.landoni@usi.ch
Università della Svizzera Italiana
Switzerland

Maria Soledad Pera
m.s.pera@tudelft.nl
Web Information Systems - TU Delft
The Netherlands

ABSTRACT

The influence of ChatGPT and similar models on education is being increasingly discussed. With the current level of enthusiasm among users, ChatGPT is envisioned as having great potential. As generative models are unpredictable in terms of producing biased, harmful, and unsafe content, we argue that they should be comprehensively tested for more vulnerable groups, such as children, to understand what role they can play and what training and supervision are necessary. Here, we present the results of a preliminary exploration aiming to understand whether ChatGPT can adapt to support *children* in completing *information discovery tasks* in the education context. We analyze ChatGPT responses to search prompts related to the 4^{th} grade classroom curriculum using a variety of lenses (e.g., readability and language) to identify open challenges and limitations that must be addressed by interdisciplinary communities.

ACM Reference Format:
Emiliana Murgia, Zahra Abbasiantaeb, Mohammad Aliannejadi, Theo Huibers, Monica Landoni, and Maria Soledad Pera. 2023. ChatGPT in the Classroom: A Preliminary Exploration on the Feasibility of Adapting ChatGPT to Support Children's Information Discovery. In *UMAP '23 Adjunct: Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23 Adjunct), June 26–29, 2023, Limassol, Cyprus.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3563359.3597399

1 INTRODUCTION

We are witnessing the ever-growing development and interest in large language models[1] (LLMs) and how they can be used to address information-seeking tasks. From designing AI-powered bots such as ChatGPT [4] and YouChat [3], to model-based retrieval models [34], generative AI has attracted increasing attention in the community.

ChatGPT is here; researchers, practitioners, and everyday users alike are aware of it, yet all are still trying to understand the audiences, tasks, and contexts in which ChatGPT may be useful [38]. Emerging technologies like ChatGPT provide "opportunities for an active and meaningful learning environment in the school context, provoking important reflections on what is expected from the 21st-century school" [40]. Indeed, early adopters in this context have begun integrating ChatGPT in their schools, focusing on its potential benefits to improve teaching as well as personalizing the learning experience [25]. At the same time, there are already concerns about the instrument itself and its likely misuse [5]. Moreover, model bias [33, 43] and hallucination [24] are two well-known problems of generative models which can affect all groups of users but would be harder for young users to detect and combat them [30, 35, 37].

The educational context encompasses many teaching and learning tasks that AI technologies could enable. It serves a wide range of individuals; from educators themselves to students of all ages. We argue that with the rapid and ever-changing landscape of AI technologies for the educational context, it is critical to identify *how to explore and assess the impact that AI technologies can have in the educational context?* and also *what are the advantages and inevitable challenges ahead?* To start answering these questions, we conduct a set of preliminary *quantitative and qualitative explorations* on the use of ChatGPT. To control the scope of our work, we follow the framework introduced by Landoni et al. [26], which guides the design and assessment of information retrieval technology through four pillars. In our case, these pillars are children aged 10–11 (4^{th} grade in primary school) as the *user group*, classrooms as the *environment*, information discovery as the high level *task*, and information produced by ChatGPT as the *strategy*. Specifically, we reflect on ChatGPT's ability to adapt to serve a particular user group and context by examining responses generated by ChatGPT for a number of prompts common to the 4^{th} grade history curriculum from different lenses, including readability, language, and type of tasks.

Findings reveal that ChatGPT could support children's information discovery by adapting the language literacy level of the

*All authors contributed equally to this research.

[1]Large language models, like GPT-3, have the ability to perform tasks for which they were never explicitly trained for given human language descriptions or examples [13, 32], i.e., these models can be adapted to accomplish specific tasks via prompts [45].

UMAP '23 Adjunct, June 26–29, 2023, Limassol, Cyprus

ACM ISBN 978-1-4503-9891-6/23/06.
https://doi.org/10.1145/3563359.3597399

answers even if the formal assessment of readability–in agreement with direct feedback from children–shows that responses are more complex than required by this age group. Information pollution [27], however, can be so naturally blended into ChatGPT responses that it is difficult for children to spot. Hence the need to train teachers and children to be critical of ChatGPT responses and verify sources. For some areas, like safety-related issues, there can be "significant risks when using ChatGPT as a source of information and advice" [38]. With this work, we build on this discourse by exploring the risk, but more importantly the opportunities, of using ChatGPT to aid the classroom context.

2 EXPERIMENTAL SETUP

Prompt design. To gather data for analysis, we adopt the prompts defined by expert educators to guide the completion of an online inquiry assignment in the 4^{th} grade [7]. Specifically, we rely on twelve questions related to Ancient Rome, a history topic common in the school curriculum (P_{ID}). As generative LLMs are affected by '*hallucinations*', i.e., they are "prone to hallucinate unintended text, which degrades the system performance and fails to meet user expectations in many real-world scenarios" [24], we ask an expert educator to define three prompts aligned with Ancient Rome but referring to fictional historical events/figures (P_H) to enable appraisal of ChatGPT in this context.

Prompt categories. We categorize prompts as in [7, 26] based on the type of interactions they elicit: (i) *fact-finding* are straight–forward prompts that require a precise answer; (ii) *open-ended* are open prompts that require a short textual description as an answer; and (iii) *multi-step* are complex prompts that require connecting information to find an answer. By considering prompts referring to varied task categories (6 fact-finding, 4 open-ended, and 2 multi-step, respectively in P_{ID}, uniformly distributed for P_H), we can examine ChatGPT's behavior when faced with prompts of increased complexity (see Table 1).

Language. To examine variability in ChatGPT's performance due to language, we turn to native speakers to translate P_{ID} and P_H from their original Italian to English.

Data collection. We use two strategies: (i) $ChatGPT_D$ where we elicit ChatGPT responses for P_{ID} and P_H; and (ii) $ChatGPT_{CF}$ where we add the phrase "explained to a fourth grader" (in the respective language) to P_{ID} and P_H. We posit that including an explicit target audience could yield child-friendly responses, fitting the target audience under study. This results in 60 text samples, i.e., responses, uniformly distributed across language and collection strategy.

Exploration We probe the responses generated by ChatGPT using eight measures that capture the *linguistic and stylistic complexity* of ChatGPT responses. For this, we use Python's `textstat` [2] and `textcomplexity` [1], which provide options for scrutinizing texts in Italian and English. In particular, we analyze the generated results in terms of (i) word count; (ii) unique word count; (iii) sentence cont; (iv) average sentence length; (v) Flesh Reading Ease [19]; (vi) reading time [15]; (vii) entropy [42]; and (viii) closeness [21]. When juxtaposing results across ChatGPT strategies, prompt type, i.e., P_{ID} and P_H, prompt categorization, and language, we determine significance using t-test, $p < 0.05$.

We also gauge the *suitability* of ChatGPT responses for the main stakeholders in the educational context under study: children. With required ethical considerations accounted for[2], we turn to 55 students in the 4^{th} grade of a primary school in Italy. We share each response generated by $ChatGPT_{CF}$ for P_{ID} and ask them to "Rate the readability of this text." For feedback elicitation, we use an adapted version of the popular 5-point Likert scale, where 1 indicates very difficult to comprehend and 5 very easy, based on emojis–a common practice when involving young users [41].

3 RESULTS, DISCUSSION, AND IMPLICATIONS

Here, we present the results of our initial exploration of ChatGPT.

Can ChatGPT adapt its responses to primary school students? We examine differences in linguistic and stylistic complexity measures inferred from responses generated by $ChatGPT_D$ and $ChatGPT_{CF}$ for P_{ID} in their original Italian. As shown in Fig. 1a, the average number of words significantly decreases for $ChatGPT_{CF}$ responses compared to $ChatGPT_D$. The same is true for the average sentence length and average reading time (the estimated amount of time it takes to read a given text). At the same time, the average number of sentences per response significantly increases for $ChatGPT_{CF}$, when compared to $ChatGPT_D$; we see this as an expected trade-off to produce shorter, easier-to-read sentences. On average, the number of unique words is comparable across strategies. Yet, it emerges from Fig. 1a that $ChatGPT_D$ responses result in a wider range of unique words.

Flesh Reading Ease scores (computed using the formula specifically adapted for Italian text [20]), which determine the degree of difficulty of text samples, significantly decrease from an average score of 70 to less than 60 for responses generated using $ChatGPT_{CF}$ and $ChatGPT_D$, respectively. Scores of 70 and above indicate fairly easy-to-read text, whereas scores of '60-69' and '50-59' signal standard and fairly difficult-to-read texts, respectively. The mean entropy of $ChatGPT_{CF}$ responses is lower than $ChatGPT_D$ (Fig. 2a). This difference indicates that $ChatGPT_{CF}$ is more likely to predict correct terms and is, therefore, more certain than $ChatGPT_D$ in generating responses. In addition, the mean closeness of responses generated by $ChatGPT_{CF}$ is higher than $ChatGPT_D$. The closeness metric is inversely correlated to the average length of the shortest path between nodes of the dependency tree. Hence, the generated responses for children are less complex than $ChatGPT_D$ in terms of dependency between terms in a sentence.

From these results, we infer that when explicitly specifying the target audience, ChatGPT adapts its responses to produce easier-to-decode [14] text with a more limited vocabulary and shorter and simpler sentences. Outcomes match those reported by Benzon [11] who states that "ChatGPT can adjust its level of discourse to accommodate children of various ages." Note, however, that text samples scored in the '70-79' range of Flesh Reading Ease reflect material suitable for 7^{th} graders (i.e., 13 to 14-year-olds). This is a limitation, as responses are meant to match the abilities of 4^{th} graders.

Can ChatGPT support different types of primary school inquiry tasks? We assess ChatGPT's versatility when addressing

[2]The study was part of regular school activities, and ran on a voluntary basis with consent from the principal and teachers from the host Institution.

Table 1: Sample prompts (translated from Italian).

ID	Question	Category
1	Why did the first Romans settle on the hills?	Open ended (in P_{ID})
2	Were the kings of Rome chosen by birth or election?	Fact-finding (in P_{ID})
3	How long did the monarchy last?	Multi-step (in P_{ID})
4	Who was King Tarquinius the Pisquano?	Fact-finding (in P_H)

prompts for inquiry tasks of increased levels of complexity. From Fig. 1c and 2c, we detect that except for entropy and closeness, trends reported thus far are not consistent across categories for P_{ID}. They align with those observed for fact-finding prompts but seldom coincide with those emerging from open-ended and multi-step prompts. We attribute this to the nature of the tasks and their growing complexity. Still, it is clear that ChatGPT needs to adjust to better enable the types of inquiry tasks that are common to the educational context.

Can ChatGPT alter its reactions to fictional prompts? To scrutinize whether ChatGPT produces responses of different styles when it comes to prompts referring to fictional historical figures and events related to Ancient Rome, we compare linguistic and stylistic complexity measures computed for P_{ID} vs. P_H. It is visible from Fig. 1a that scores computed for P_{ID} rarely match those for P_H. Among salient differences, we highlight a significant increase in average sentence length as well as a decrease in the average number of sentences and average reading time for response generated by $ChatGPT_{CF}$ for P_{ID} when compared to P_H.

Manually inspecting the samples, we see an interesting pattern among responses to P_H: for 2 out of the 3 prompts, $ChatGPT_{CF}$ seems to presume that the user made a typo on the historical figure/event mentioned in the prompt (so did $ChatGPT_D$). Accordingly, its responses do not address the intent of the prompt. On the remaining prompt, $ChatGPT_{CF}$ states that it does not recognize the existence of that historical event, and proceeds to discuss a similarly-named event. These findings further showcase issues of *hallucinations* impacting generative LLMs. We argue that the fact that ChatGPT does not modify its behavior when responding to fictional prompts is a particular concern for the audience of this study, who seldom question the veracity of the online information presented to them [30]. As ChatGPT uses natural language, children may be prone (even more so than when using search engines) to believe the content produced is reliable. This showcases the need to train children on how to engage with ChatGPT and to be critical in their judgements.

Are ChatGPT responses useful to primary school students? We examine the feedback elicited from 4^{th} grade students regarding their ability to understand $ChatGPT_{CF}$-generated text–a proxy that enables us to judge the perceived fit of ChatGPT to support young users who need to be able to comprehend responses produced for them to be of use.

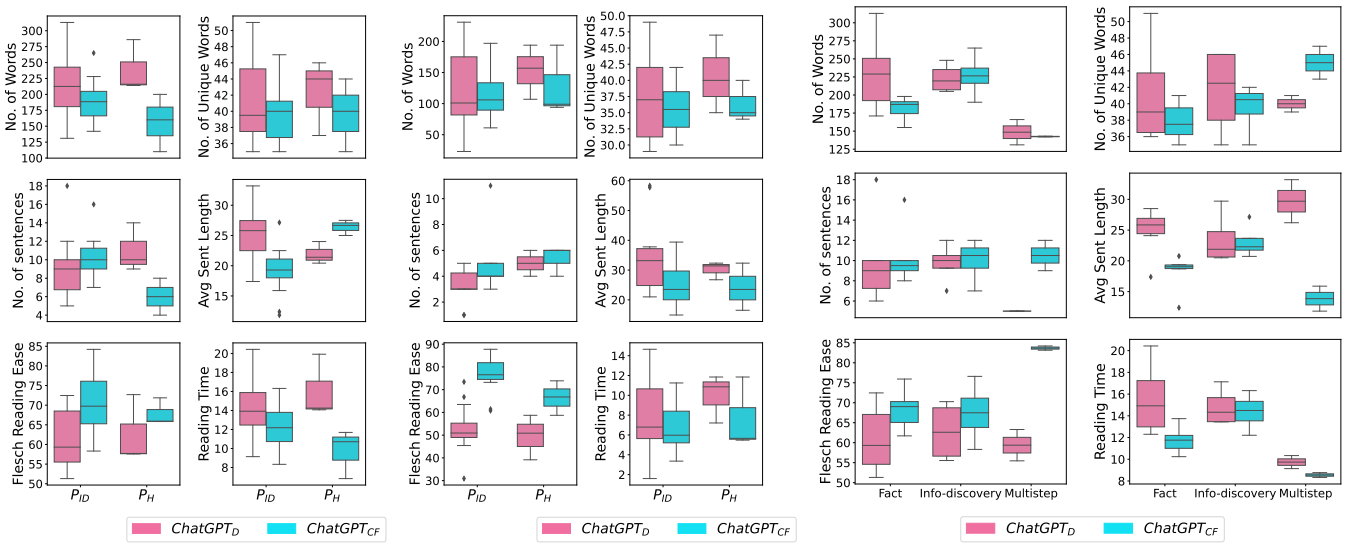
The distribution captured in Fig. 3 points to children placing the readability of most prompts between neutral and good (i.e., Neutral and Happy in terms of emojis). The responses to the two prompts deemed the most readable (5 and 10) were relatively short and in the case of prompt 10, the response was a short story about a legend in ancient Rome, which children are used to and contained less specific lexicon. Agreeing with the discussion on readability scores presented earlier, the samples produced by $ChatGPT_{CF}$ appear too complex for a complete understanding by a 4^{th} grader. Overall, we surmise that children understood the samples presented to them, but were not completely satisfied; suggesting in turn that ChatGPT is not quite ready to really help children[3].

Does language affect ChatGPT's ability to adapt its responses to primary school students? For tasks like sentence understanding, ChatGPT fares better for English-written text, as opposed to other languages [10]. This is why we look for possible discrepancies in scores estimated from responses for P_{ID} in Italian vs. English. Contrasting Fig. 1a and 2a with Fig. 1b and 2b, we notice how trends and significant differences in scores observed for $ChatGPT_D$ vs. $ChatGPT_{CF}$ remain the same regardless of the language.

As seen from the analysis of the responses for P_{ID} in Italian, those produced using $ChatGPT_{CF}$ for prompts written in English yield significantly higher Flesh Reading Ease scores i.e., easier to read, than those using $ChatGPT_D$ (computed using the corresponding Flesh Reading Ease formula, depending on the language of the text sample examined). The average score of responses produced using $ChatGPT_D$ are closer to 50, i.e., 'fairly difficult' to read, whereas those generated by $ChatGPT_{CF}$ are closer to 75 (and above), i.e., 'fairly easy.' The average reading time significantly increases (significantly decreases, resp.) for responses produced by $ChatGPT_D$ ($ChatGPT_{CF}$, resp.) for English prompts with respect to their Italian counterparts. These results indicate that, while still above the skills of 4^{th} graders, ChatGPT is more likely to provide simpler text in English than in Italian. This is anticipated, given ChatGPT's self-proclaimed preference for English (Fig. 4). Regarding closeness, the generated responses in English follow the same trend as in Italian; with the difference between the closeness of the responses generated by $ChatGPT_{CF}$ and $ChatGPT_D$ being more prominent in English. Entropy is a measure of uncertainty that means lower entropy indicates more certainty than a higher value of entropy. For both $ChatGPT_D$ and $ChatGPT_{CF}$ the mean entropy of responses for P_{ID} is lower than for P_H, which indicates that chatGPT is more certain in generating the response for P_{ID} than P_H, a promising feature when considering how important it is to prevent children's exposure to information pollution [27].

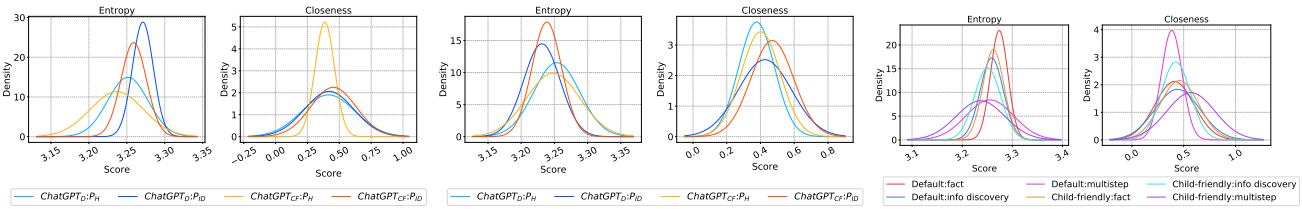
Could ChatGPT replace web search in the primary school environment? Haque et al. [22] question whether ChatGPT could replace search engines, as it presents "information conveniently by selecting the most appropriate information and explaining it in simple terms." Neither ChatGPT nor web search engines were designed specifically for children or the educational context. Web search engines, however, are the go-to portals to resources that

[3]For a more detailed analysis of children's perception of ChatGPT's ability to produce easy to read and comprehend text for the classroom, see [36].



(a) Italian (b) English (c) Italian by task category

Figure 1: Overview of text-based measures computed on responses generated by ChatGPT using the default ($ChatGPT_D$) vs. child-friendly ($ChatGPT_{CF}$) strategies for real (P_{ID}) and fictional (P_H) prompts



(a) Italian (b) English (c) Italian by task category

Figure 2: Closeness and Entropy of ChatGPT responses.

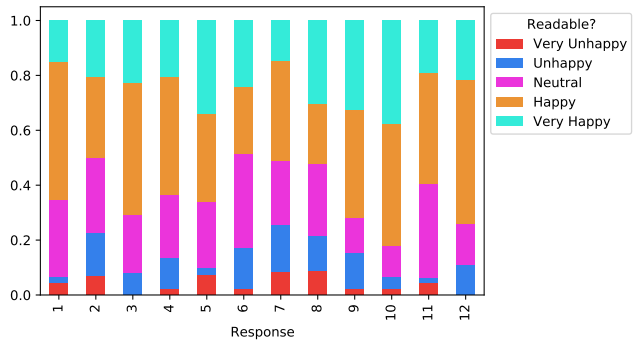


Figure 3: Children's perceptions on the ease of comprehension of $ChatGPT_{CF}$ responses to P_{ID}.

can enable teaching and learning in formal and informal settings [9, 18, 29, 31, 44]. They are embedded in the educational context, even when their limitations to retrieve and prioritize educational resources are well documented [8, 39]. As a generative model, ChatGPT may not always produce reliable responses, in turn exposing children to information pollution. It still offers well-written and easily understandable answers to a wide range of prompts. It is natural then to question whether ChatGPT can be used to help children with their online inquiries, as it can alleviate some common challenges they face, such as formulating effective search queries and finding relevant resources from the search engine results pages (SERP) when completing educational tasks [9].

Informed by the results discussed thus far, ChatGPT could ease query formulation: directly using assignment prompts, students can access complete answers. This could be to the detriment of developing a skill–query formulation–required in the digital ecosystem we inhabit. ChatGPT also removes the need for SERP exploration by providing direct responses and hiding sources under a smoothly written piece of text ready to be used, with no more access to indirect clues of their quality as found in SERP. With children seldom questioning source reliability [30], what are the consequences of ChatGPT hallucinations? Web search engines and ChatGPT are

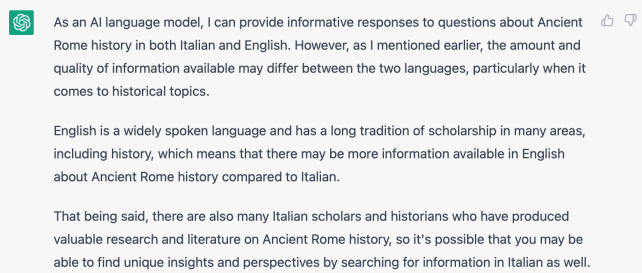


Figure 4: ChatGPT and its preference for English.

examples of technologies made for general audiences that can therefore make generalizations and misinterpret users' needs [16, 17]. In this case, what are the implications for the educational context when ChatGPT generates incorrect responses that do not necessarily match the intent expressed in the prompt used to elicit a response? Presenting younger user groups with material suitable for their skills is a challenge for web search engines and ChatGPT. The average readability level of resources retrieved in response to children's queries is significantly above what they can understand [8, 12]. The same is true for ChatGPT. Recall that the estimated readability of the responses for P_{ID} reflects the skills of 7^{th} graders. Worth mentioning, however, is the fact that in the case of search engines, children can browse and actively select retrieved documents based on their own judgments of readability; they are not presented with choices regarding text complexity when dealing with ChatGPT unless they engage in prompt engineering [45].

Query formulation impacts retrieval effectiveness and the retrieval of different types of resources [6]. We question if variations on prompts used to elicit responses from ChatGPT would result in more (or less) effective responses. A preliminary manual examination of responses produced by modifying the phrase used in $ChatGPT_{CF}$ to elicit child-friendly outcomes indicates that, much like query variations on web search engines, prompt variations influence outcomes. E.g., when asked for advice for a 4^{th} grader on how to prepare a presentation on a school topic about who were the inhabitants of the area where Rome was built (a variation of one of the prompts in P_{ID}), ChatGPT responded by suggesting how to conduct an online search to gather information, it pointed out online sources that could provide suitable content, and it encouraged using text and images to create engaging presentations. In this case, in lieu of a precise answer, ChatGPT offered scaffolding on how to approach information discovery for the classroom; evincing behavior closer to that of a potential educational agent [28].

In the end, generative LLMs and search engines could trigger different roles (i.e., passive vs. active) in children engaging with the technology itself and critically assessing the offered results. Consequently, both could assist children with search activities in the classroom based on how proficient they already are.

4 CONCLUDING REMARKS

AI technologies in vogue nowadays, like OpenAI's ChatGPT, Google's Bard, and Bing's AI Chat, which are "trained on unprecedented amounts of data and able to engage in astonishingly diverse conversations" [23], are already used by millions. Yet, we know little about their adaptability and applicability for specific contexts or their impact on user groups for which they were not explicitly designed. With this work, we aimed to bring attention to the use of AI in education by children. We discussed lessons learned from a preliminary exploration of the extent to which ChatGPT can adapt to support primary school inquiry assignments. We considered different lenses to scrutinize ChatGPT's feasibility to adapt to support young users (e.g., user group, inquiry task type, and language). We used a combination of quantitative and qualitative data to support discussion; we also involved members of the target community to assess the usefulness of responses to prompts derived from search tasks set by teachers. This approach successfully guided our exploration and grounded it in the classroom.

While limited by the number of prompts, this work showcases the need for further investigations to understand the potential socio-technical implications inherent to the use of generative LLMs in the educational context. We did not repeat response generation multiple times for the same prompt. Yet, there are already indications of how models like ChatGPT dynamically change as they are used. Further, prompts can impact generated responses [45]. We plan to consider this in future iterations of our work. Other extensions to this work include probing performance on other topics common to the primary school curriculum and languages, i.e., beyond history and Italian.

ChatGPT has managed to attract a lot of attention, concerns, and even anxiety from educators. We share a more objective account of its potential and limits with the community. Aware of the fact that this technology is here to stay, it is worth getting a better understanding of how it can support education–facilitate teaching while enabling personalized learning–and at the same time how to deal with its shortcomings, in order to set the right expectations for all involved stakeholders, starting from children.

REFERENCES

[1] 2022. TextComplexity. https://pypi.org/project/textcomplexity/
[2] 2022. Textstat. https://pypi.org/project/textstat/
[3] 2023. The Ai Search Engine you control | AI chat & apps. https://you.com/
[4] 2023. Introducing ChatGPT. https://openai.com/blog/chatgpt
[5] 2023. World Economic Forum on Instagram: "is the pen mightier than the AI-powered chatbot? learn more about chatbots by tapping on the link in our bio. follow our annual meeting at Davos from 16 - 20 Jan". https://www.instagram.com/p/CnboSxKoNao/?igshid=YmMyMTA2M2Y
[6] Marwah Alaofi, Luke Gallagher, Dana McKay, Lauren L Saling, Mark Sanderson, Falk Scholer, Damiano Spina, and Ryen W White. 2022. Where Do Queries Come From?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2850–2862.
[7] Mohammad Aliannejadi, Monica Landoni, Theo Huibers, Emiliana Murgia, and Maria Soledad Pera. 2021. Children's Perspective on How Emojis Help Them to Recognise Relevant Results: Do Actions Speak Louder Than Words?. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 301–305.
[8] Oghenemaro Anuyah, Ashlee Milton, Michael Green, and Maria Soledad Pera. 2020. An empirical analysis of search engines' response to web search queries associated with the classroom setting. *Aslib Journal of Information Management* 72, 1 (2020), 88–111.
[9] Ion Madrazo Azpiazu, Nevena Dragovic, Maria Soledad Pera, and Jerry Alan Fails. 2017. Online searching and learning: YUM and other search tools for children and teachers. *Information Retrieval Journal* 20 (2017), 524–545.
[10] 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* (2023).

[11] William L Benzon. 2023. Discursive Competence in ChatGPT, Part 1: Talking with Dragons. (2023). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4318832

[12] Dania Bilal and Li-Min Huang. 2019. Readability and word complexity of serps snippets and web pages on children's search queries: Google vs bing. *Aslib Journal of Information Management* (2019).

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

[14] Donald L Compton, Amanda C Appleton, and Michelle K Hosp. 2004. Exploring the relationship between text-leveling systems and reading accuracy and fluency in second-grade students who are average and poor decoders. *Learning Disabilities Research & Practice* 19, 3 (2004), 176–184.

[15] Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109, 2 (2008), 193–210.

[16] Brody Downs, Maria Soledad Pera, Katherine Landau Wright, Casey Kennington, and Jerry Alan Fails. 2022. KidSpell: Making a difference in spellchecking for children. *International Journal of Child-Computer Interaction* 32 (2022), 100373.

[17] Nevena Dragovic, Ion Madrazo Azpiazu, and Maria Soledad Pera. 2016. " Is Sven Seven?" A Search Intent Module for Children. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 885–888.

[18] Michael D Ekstrand, Katherine Landau Wright, and Maria Soledad Pera. 2020. Enhancing classroom instruction with online news. *Aslib Journal of Information Management* 72, 5 (2020), 725–744.

[19] James N Farr, James J Jenkins, and Donald G Paterson. 1951. Simplification of Flesch reading ease formula. *Journal of applied psychology* 35, 5 (1951), 333.

[20] Valerio Franchina and Roberto Vacca. 1986. Adaptation of Flesh readability index on a bilingual text written by the same author both in Italian and English languages. *Linguaggi* 3 (1986), 47–49.

[21] Linton C. Freeman. 1978. Centrality in social networks conceptual clarification. *Social Networks* 1, 3 (1978), 215–239.

[22] Mubin Ul Haque, Isuru Dharmadasa, Zarrin Tasnim Sworna, Roshan Namal Rajapakse, and Hussain Ahmad. 2022. " I think this is the most disruptive technology": Exploring Sentiments of ChatGPT Early Adopters using Twitter Data. *arXiv preprint arXiv:2212.05856* (2022).

[23] Natali Helberger and Nicholas Diakopoulos. 2023. Chatgpt and the AI act. https://policyreview.info/essay/chatgpt-and-ai-act

[24] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *Comput. Surveys* (2022).

[25] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. (2023).

[26] Monica Landoni, Davide Matteri, Emiliana Murgia, Theo Huibers, and Maria Soledad Pera. 2019. Sonny, Cerca! evaluating the impact of using a vocal assistant to search at school. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*. Springer, 101–113.

[27] Monica Landoni, Emiliana Murgia, Theo Huibers, and Maria Soledad Pera. 2023. How does Information Pollution Challenge Children's Right to Information Access?. In *3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval co-located with ECIR '23. CEUR Workshop Proceedings*, Vol. 3359. CEUR-WS, 250–253.

[28] Monica Landoni, Maria Soledad Pera, Emiliana Murgia, and Theo Huibers. 2022. Let's Learn from Children: Scaffolding to Enable Search as Learning in the Educational Environment. *arXiv preprint arXiv:2209.02338* (2022).

[29] Konstantinos Lavidas, Anthi Achriani, Stavros Athanassopoulos, Ioannis Messinis, and Sotiris Kotsiantis. 2020. University students' intention to use search engines for research purposes: A structural equation modeling approach. *Education and Information Technologies* 25 (2020), 2463–2479.

[30] Eugène Loos, Loredana Ivan, and Donald Leu. 2018. "Save the Pacific Northwest tree octopus": a hoax revisited. Or: How vulnerable are school children to fake news? *Information and Learning Science* (2018).

[31] Silvia B Lovato, Anne Marie Piper, and Ellen A Wartella. 2019. Hey Google, do unicorns exist? Conversational agents as a path to answers to children's questions. In *Proceedings of the 18th ACM international conference on interaction design and children*. 301–313.

[32] Christopher D Manning. 2022. Human language understanding & reasoning. *Daedalus* 151, 2 (2022), 127–138.

[33] Justus Mattern, Zhijing Jin, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Understanding Stereotypes in Language Models: Towards Robust Measurement and Zero-Shot Debiasing. *arXiv preprint arXiv:2212.10678* (2022).

[34] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: making domain experts out of dilettantes. *SIGIR Forum* 55, 1 (2021), 13:1–13:27.

[35] Marisa Meyer, Victoria Adkins, Nalingna Yuan, Heidi M Weeks, Yung-Ju Chang, and Jenny Radesky. 2019. Advertising in young children's apps: A content analysis. *Journal of developmental & behavioral pediatrics* 40, 1 (2019), 32–39.

[36] Emiliana Murgia, Maria Soledad Pera, Theo Huibers, and Monica Landoni. 2023. Children on ChatGPT Readability in an Educational Context: Myth or Opportunity?. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23 Adjunct)*. 8 pages.

[37] Grace W Murray. 2021. Who is more trustworthy, Alexa or Mom?: Children's selective trust in a digital age. (2021).

[38] Oscar Oviedo-Trespalacios, Amy E. Peden, Thomas Cole-Hunter, Arianna Costantini, Milad Haghani, J.E. Rod., Sage Kelly, Helma Torkamaan, Amina Tariq, James David Albert Newton, and et al. 2023. The risks of using CHATGPT to obtain common safety-related information and advice. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4346827

[39] Jodi Pilgrim. 2019. Are we preparing students for the web in the wild? An analysis of features of websites for children. *The Journal of Literacy and Technology* 20, 2 (2019), 97–124.

[40] Charles Pimentel. 2022. Is ChatGPT a threat to education? For banking model of education, yes. (2022). https://fellows.fablearn.org/blogs/

[41] Janet C Read and Stuart MacFarlane. 2006. Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In *Proceedings of the 2006 conference on Interaction design and children*. 81–88.

[42] Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.

[43] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3407–3412.

[44] Hamid Slimani, Oussama Hamal, Nour-Eddine El Faddouli, Samir Bennani, and Naila Amrous. 2020. The hybrid recommendation of digital educational resources in a distance learning environment: The case of MOOC. In *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*. 1–9.

[45] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382* (2023).