EXPLORING LARGE LANGUAGE MODELS FOR THE EDUCATION OF INDIVIDUALS WITH COGNITIVE IMPAIRMENTS

A. Fiora, F. Piferi, P. Crovari, F. Garzotto

I3lab - Innovative Interactive Interfaces Laboratory
Department of Electronics, Information, and Bioengineering
Politecnico di Milano (ITALY)

Abstract

We investigate the application of Large Language Models (LLMs) in educational contexts for individuals with Cognitive Disorders (CD) by evaluating the ability of these models to generate accessible and engaging content tailored to the needs of this specific group. The research adopts a comparative analysis approach, using guidelines formulated with expert input to assess the ability of the LLMs to adapt to an accessible communication. The findings reveal varying capabilities among different LLMs in adhering to these guidelines, highlighting the adaptability of AI tools for special education. This study underscores the potential of LLMs to revolutionize learning experiences for individuals with CD, advocating for continued exploration and enhancement in this area to develop more inclusive and effective educational technologies.

Keywords: Large Language Model, Special Education, Cognitive Disorders

1 INTRODUCTION

Generative Artificial Intelligence (Gen AI) is artificial intelligence capable of generating text, images, or other media, using generative models capable of learning the patterns and structure of their input training data and then generating new data with similar characteristics [1].

Large Language Models (LLMs) are a type of Generative AI that can comprehend and generate human-like text. LLMs are trained on an unprecedented amount and variety of text, including hundreds of billions of utterances from online digital sources, such as web pages, books, magazines, social media posts, Wikipedia, and online forums [2], [3], [4]. Thanks to better modeling and data, LLMs can measure subtle, context-dependent expressions of mental states, generate human-like text to answer questions and provide information, synthesize content from a wide range of sources, engage in natural dialogue, and simulate different linguistic styles and personas [1].

LLMs hold immense potential across various domains. An increasing body of research explores the use of LLMs in education at all stages [5], [3], [4], highlighting the potential benefits of these technologies from both the learners' and the teachers' perspectives. For learners, LLMs can assist in the development of scientific and literacy skills (question-asking, reading, writing, language capability), foster creativity, support critical thinking, aid in research, information analysis, and creation of personalized practice materials (e.g., summaries and explanations), which can help improve student performance and contribute to enrich and diversify the learning experience [2], [3], [4]. They can also support language learning by offering interactive conversations, vocabulary explanations, and grammar assistance [1], [6]. LLMs can support the activities of educators and decrease the teaching workload, facilitating the design and creation of educational content, lesson plans, and assessment items to evaluate student work and provide feedback.

While LLMs offer valuable support to education, many authors pinpoint the need to acknowledge their weaknesses and uncertainties, which may prevent LLMs' successful and responsible integration into learning and teaching processes [2], [3], [4]. Such limitations include biases in the generated content, unexpected brittleness in relatively simple tasks, difficulty adapting to the diverse needs of students and teachers and providing the level of personalization required for effective learning, lack of explainability and appropriate interfaces for an effective interaction between learners/teachers and the generative tools.

A notable gap in the current state of the art is the limited investigation of LLMs in the educational processes for people with disabilities, particularly those with Cognitive Disorders (CD). CD is characterized by impairments in different abilities such as memory, perception, information processing, problem-solving language, and social cognition.

Verma et al. [7] pinpoint the potential of LLM to empower this population. The research employed a mixed-methods approach, involving an evaluation of one of the most popular LLM (ChatGPT) to assess its utility for individuals with intellectual disabilities and a survey with 17 neurodiverse participants. The results indicate that ChatGPT shows promise in bridging the digital divide for this population and emphasize the importance of enriching the interaction and interface features to improve accessibility.

Ciampa et al. [2] argue that students with reading and writing impairments (e.g., students with dyslexia) can benefit from LLMs as an assistive technology tool that facilitates them to access textual materials, assist them with spelling, grammar, and idea generation, and reduces frustration and embarrassment during the writing process. Individuals with Attention Deficit Hyperactivity Disorder (ADHD) can find support with organization and focus, to initiate writing, structure thoughts, and outline complex writing tasks [8].

The study reported in [9] explores the fairness of LLMs for people with disabilities and indicates the existence of biases against disabled communities in the generated content.

Our research addresses LLMs as educational assistants for people with CD. We want to understand the extent to which an LLM can accommodate the needs of such people and have a foundation to decide if it's feasible to use such models for generating content.

We define as *accessible* the content that is appropriate from a communication perspective for people with CD. LLMs, by default, generate a human-like text with all the traits of an average, well-educated person, embodying all the characteristics, defects, biases, and vices of such an individual [9]. This can be problematic when dealing with people with CD, who are known to adapt less to common speech and communication patterns and to understand to a smaller extent a text that follows such patterns. For this reason, we intend to answer the two following research questions:

RQ1: Do the available LLMs produce content that is "accessible" for people with CD?

RQ2: Can information requests incorporating accessibility guidelines make the generated content more appropriate for the needs of persons with CD?

We investigate aspects that are a *prerequisite* for making LLMs helpful for this population but are independent of the specific learning goal or task for which this technology is used. Such aspects are related to the *accessibility* of the *dialogic interaction between learners and LLMs*. Specifically, we focus on HOW something is said by an LLM (not on WHAT is said) and whether the generated content is *appropriate from* a *communication perspective* for the needs of a person with CD.

To answer our research question, we constructed and carried out a comparative analysis, evaluating both the standard behavior of the available LLMs and their behavior when guided through the techniques of prompt engineering. We then analyzed the results to answer the two research questions.

Our work brings three main contributions:

1. A set of guidelines to generate accessible text with Generative AI for people with CD and their translation into prompts to be given to an LLM;
2. A testing methodology to assess the accessibility for people with CD of the text produced by an LLM;
3. A comparative analysis of the main commercially available models in their ability to produce accessible text for people with cognitive disorders.

Our results lay the foundation for designing educational tools based on LLMs suitable for people with CD, paving the way for a whole new generation of educational technologies.

2 METHODOLOGY

To address the research questions, we performed a comparative analysis of Multiple LLMs: ChatGPT3.5 and ChatGPT4 [6], the two versions of the LLM from OpenAI, Bard [10] and Palm2 [11] from Google, and Llama2 [12] from Meta. To run the evaluation, we accessed the version of the models that was available online the week of Oct 2nd, 2023.

First, we define LLM-based generated content as *accessible* if it is *appropriate from* a *communication perspective* for the needs of a person with CD, with the objective of passing all the communicative content in its entirety, with ease, and without miscomprehensions [13].

The behavior and output of an LLM can be guided through careful crafting of input text (*prompting*) in a way that encourages the model to follow specific tasks and directives in the desired way. The quality of results depends on the quality of the requests, i.e., how much information you provide as input and how well-crafted it is [13], and, of course, on the quality and quantity of the training set, which we postulated as immutable for the purposes of this paper.

2.1 Guideline selection

We evaluated each LMM against a set of linguistic guidelines incorporating principles created with the help of expert educators and caregivers specialized in CD and tailored for the communication needs of our target population. We derived communication guidelines from best practices in the State of the Art about dealing with neurodiverse people and selected them among the most common and broad guidelines [14], [15]. These guidelines emphasize using simple and concise language, avoiding filler words, refraining from figurative language, and structuring sentences straightforwardly, which are considered best practices that help people with CD avoid distractions and misunderstanding while reading text or interacting with people. The final 9 guidelines are shown in the first column of Table 1.

Table 1. Guidelines and related prompts (see §2.4)

	Guideline	***Prompt for the LLM***
G1	Use of simple words [14]	Use simple words.
G2	Avoidance of filler words [15]	Write as long as you normally would, but do not use filler words (filler words are words that enhance the flow of the text but do not add information to the text).
G3	Avoidance of figurative language [14]	Do not use figurative language.
G4	Use of short and succinct language [15]	Use only short and succinct phrases.
G5	Use of graphical aids to create a clearer text division (paragraphs, numbered lists, titles, sections…) [14]	Use a clear division of the text.
G6	Use of only sentences in the simple form SVO (subject-verb-object, without chains of indirect complements) [14]	Use only sentences structured in a subject-verb-object way.
G7	Use a controlled number of paragraphs (according to the user's profile)	Write in *n* paragraphs.
G8	Use of a controlled number of words (according to the user's profile)	You do not have to change your answer except that each answer must count *n* words. Not a word more, not a word less, exactly n.
G9	Use of short words [15], [17]	Use words with a maximum of 7 letters each.

2.2 Topic selection and generation

We carefully considered the requests for the LLMs to create content for evaluation, aiming to neutralize – or at least balance – potential biases that might favor or disfavor a specific guideline. Consequently, we established five broad topic categories for formulating these requests, to cover most scopes in which the LLM could be asked to produce text. This was done since requests coming from different topic categories are addressed in different writing styles that make use of the same speech characteristic we are measuring in vastly different ways – for example, a story generation may use more figurative language than a technical explanation – so a broader range of topic categories averages this variety. Moreover, while each guideline was tested with unique requests to prevent a single poorly chosen question from skewing the entire assessment, consistency was maintained by testing all LLMs with the same request for each guideline and topic category. This ensures any performance differences are due to the LLMs' capabilities and not the nature of the prompts. The final requests are factual, simple, and short; they mention the topic of interest and the literary type of the output (e.g., essay, story, explanation, instruction) and do not include any instruction or communication requirements.

The final five topics we chose are:

T1 – **Scientific Explanation**: explanation of a scientific phenomenon (e.g., *How does photosynthesis work?*)

T2 – **Humanistic Essay**: description of a piece of art or literature (e.g., *Write an essay about Frida Kahlo*)

T3 – **Invented Story**: a piece of fiction (e.g., *Write a story about a king who fell into disgrace.*)

T4 – **Practical explanation**: a practical, real-life problem or task (e.g., *How do you prepare coffee?*)

T5 – **Abstract matter**: discussion on an abstract matter (e.g., *Write an essay about mutual respect*)

2.3 Metrics

Before proceeding with the evaluation, we had to define a set of metrics to objectively measure the guideline descriptions that, being written in natural language, are inherently subjective. For each guideline *Gn*, we have defined a proxy value g, called *Guideline Score,* to closely relate the adherence to the guideline to a single numerical value that quantifies the text's compliance with the guideline. We present the metrics and describe their rationale in Table 2.

Table 2. Guideline metrics and rationales

	Metric	***Rationale***
G1	$g_w = \log(f_{THE}) - \log(f_w)$ $g = \frac{\sum g_w}{\#\ total\ words}$	The simplicity of each word is calculated with this expression, where f_{THE} is the frequency of the word "the" in the English language and f_w is the frequency of the word to be evaluated, according to [16]. The frequency values have been put in a logarithmic scale since the distribution of the words in the English language is exponential (few words are very frequent, many infrequent words). The guideline score g is the average of f_w among all the words in the text.
G2	$g = \frac{\#filler\ words}{\#\ total\ words}$	The use of filler words is evaluated by counting the filler words (words that do not add anything to a phrase, either because they are superfluous or used to fill silence) and calculating their presence in percentage of the total words.
G3	$g = \frac{\#figurative\ expressions}{\#\ total\ words}$	The use of figurative language is evaluated similarly to G2, but counting figurative expressions (phrases that carry a different meaning from the semantic one).
G4	$g = \frac{\#\ words}{\#\ sentences}$	Conciseness is evaluated by the average count of words per sentence.
G5	$g = \frac{\sum points}{\#\ total\ words}$	Graphical clarity and division is scored with a point system: every graphic division entity (a paragraph, a title, a bullet list, a subsection, a bold sentence) scores one point. The points are then normalized by the number of words in the text.
G6	$g = \frac{\#SVO\ sentences}{\#\ total\ sentences}$	The Subject-Verb-Object construction is evaluated by the percentage of only SVO sentences among all the sentences in the text.
G7	$g = \frac{\#paragraph\ provided}{\#\ paragraph\ required}$	The control on the number of paragraphs is evaluated by the relative error with respect to the required number.
G8	$g = \frac{\#\ words\ provided}{\#\ words\ required}$	The control on the number of words is evaluated as in G7.
G9	$g = \frac{\#\ words\ shorter\ then\ 7\ letters}{\#\ total\ words}$	The control on the length of words is evaluated by setting 7 as the divide between "short" and "long" words [16], and counting the percentage of short words.

2.4 LLMs Evaluation

To test the LLMs, we assigned a textual request to each guideline-topic category pair (Gn, Tn).

For each guideline, we formulated a specific prompt to instruct the LLMs, as shown in Table 1. that we concatenated to the assigned textual request. For guidelines G1 to G6, and G9, we generated a second opposite version by negating the supporting verb with *"Don't"* or removing the negation when already present. For guidelines G7 and G8, we generated two versions with two different values of n.

We devised two types of tests to evaluate the two above-mentioned classes of guidelines.

For the guidelines with a meaningful opposite (G1 to G6 and G9), we tested the LLMs under three experimental conditions: a control test (CONT) requesting only the unprompted question, a positive reinforced test (POS) asking the question together with the prompt indicating to follow the guideline, and a negative reinforced test (NEG) with the prompt requesting not to follow the guideline. This generates $7 \times 5 \times 5 \times 3 = 525$ different single texts corresponding to the 525 different permutations $(Guideline, Topic, LLM, Condition)$ of test variables.

For the numeric guidelines G7 and G8, we tested the LLMs using two different values of the parameter n to be inserted in the prompt, with a total of $2 \times 5 \times 5 \times 2 = 100$ different single-generated texts corresponding to the 100 different permutations $(Guideline, Topic, LLM, n)$ of test variables.

We ultimately scored each one of these tests by calculating its corresponding Guideline Score *g*.

2.5 Final scoring

The Guideline Scores are absolute values that cannot give by themselves any idea of how compliant to the guidelines a test is, at least without having a previous analysis of texts of various degrees of compliance. This leads to two problems to solve. First, RQ1 cannot be answered solely by looking at the numbers. Secondly, in case of a POS test that is not significantly better than the CONT test, it is not possible to distinguish between the scenario where the CONT was already following the guideline, for which the prompt has little to no effect, the scenario where the LLM is not able to fulfill the guideline request because of lack of linguistic properties, and the scenario where the guideline request is misunderstood. In principle, it is impossible to answer RQ1 and RQ2 by only looking at the absolute values or the delta between CONT and POS.

Considering all this, it becomes essential to use the NEG tests, from which we define the Prompt Efficacy Score *(PES)*, a value accounting for all the scenarios described above.

The PES is computed starting from the three guideline scores: g_{CONT}, g_{POS} and g_{NEG}, relative to the three tests for each pair (Gn, Tn) and defined as:

$$PES_{1\ldots6,9} = \frac{g_{POS} - g_{CONT}}{g_{POS}} \cdot \frac{g_{CONT} - g_{NEG}}{g_{CONT} + g_{NEG}}$$

where g_{CONT} represents the guideline score of the control text, g_{POS} is the guideline score of the positive text, and g_{NEG} is the guideline score of the negative text. This expression serves two purposes: the first fraction evaluates the improvement the prompt has on the guideline score of the POS text with respect to the CONT, while the second reduces the score the more the NEG text has a similar score to the CONT one. This way, the three scenarios mentioned above are distinguished and valued differently in this fashion: if POS is similar to CONT because CONT is already following the guidelines, it's safe to assume that if the LLM can understand the prompt, the NEG score would be very low. In this case, even if the first half of the expression is a low value, the second boosts the score upwards since it favors a very low NEG score. In any other case, the value would be near zero if no prompt has affected the scores significantly and negative if the order $g_{NEG} < g_{CONT} < g_{POS}$ is not followed.

PES expression is written to be better understood in cases where a bigger score is a better result. Still, the same formula is valid even if the opposite is true because, in that case, both numerators change sign. It must be noted that because of this, high PES can be wrongfully obtained by scores that invert the correct order – for example, when a high guideline score is better, but the scores are in the order $g_{POS} < g_{CONT} < g_{NEG}$ – however, this instance has not occurred in our testing and can be corrected with a simple check on the actual order with respect to the expected order.

As for the two guidelines G7 and G8 where CONT and NEG are not applicable, the score is a summation of individual test scores, determined by:

$$PES_{7,8} = \frac{s_i - |g_i - t_i|}{s_i}$$

where s_i defines the sensibility (the threshold below which an answer is deemed non-compliant), and $|g_i - t_i|$ denotes the absolute error between the score and the target t_i. In this case, the problems described above do not apply since the sensibility is by itself a form of grading.

Finally, we assigned to each pair $(LLM, Guideline)$ a Normalized Total PES across all topics by summing all the PES for every single topic and normalized by dividing this sum by the best possible result, calculated by taking the best scoring test of each guideline among all the LLMs and conditions. The final scores are then values ranging in the interval $[-1, 1]$, where 1 means the LLM consistently gets the best result in every test with the same rate of improvement among tests and different LLMs.

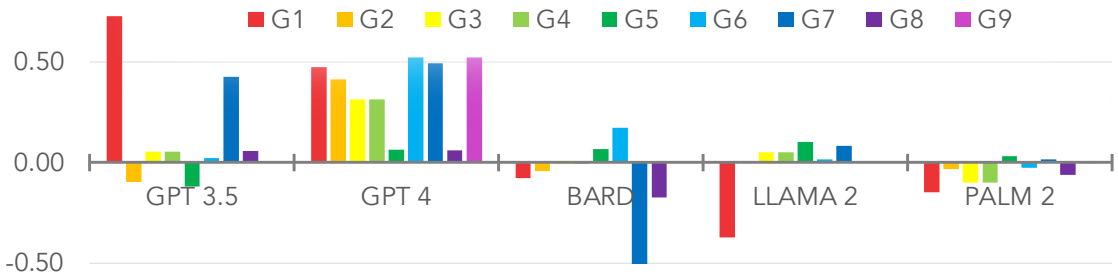
3 RESULTS

The analysis of the scores of the control tests did not highlight substantial differences among the different LLMs and a wide degree of homogeneity in the values of all the considered variables. This is to be expected since we know that LLMs mimic the language skills of the general population, which we know is not generally compliant with the guidelines.

In contrast, notable differences emerge when we consider the tests in which the guideline prompts are embedded in the request, as highlighted by Table 3 and Fig. 1, where all the Normalized Total PES are shown. For lack of space, other partial results (mainly the guideline scores and single PES) are omitted and only explained in text.

Table 3/Figure 1. Final Results – In the table is the Normalized Total PES, separated by LLM and guideline.

	GPT 3.5	***GPT 4***	***Bard***	***LLAMA 2***	***Palm 2***
G1	0.73	0.47	-0.08	-0.37	-0.15
G2	-0.10	0.41	-0.04	0.01	-0.03
G3	0.05	0.32	-0.01	0.05	-0.10
G4	0.35	0.34	0.04	0.00	0.00
G5	-0.12	0.06	0.07	0.10	0.03
G6	0.02	0.52	0.17	0.02	-0.03
G7	0.43	0.49	-0.51	0.08	0.02
G8	0.06	0.06	-0.17	0.01	-0.06
G9	0.00	0.52	0.00	0.00	0.00
Σ	**1.42**	**3.21**	**-0.53**	**-0.11**	**-0.32**



GPT-3.5 excelled in the *Simple Words* and *Short and Succinct* tests, showcasing its linguistic power. However, it lagged behind its successor GPT-4 in tasks like *No Filler Words* and *Clear Text Division*, struggling with brevity and structure. Despite these setbacks, it secured the second-highest total score of 1.42.

Dominating in rule adherence, GPT-4 outperformed others in 6 out of 9 tests, particularly in tasks requiring precise structure, like the *Number of Paragraphs* and *Simple SVO Phrases*. It uniquely varied

text sizes in the *Clear Text Division* task, reflecting its adaptability and complexity. With positive scores across all tests and an impressive total of 3.21, GPT-4 gets to a clear first place.

Bard's performance was uneven; even though it proved promising in some tasks, others were executed disastrously. It excelled in conciseness but struggled with simplicity and brevity, scoring poorly in the *Simple Words* and *No Filler Words* tests. Despite some text-structuring strengths, it scored the lowest overall at -0.53, revealing a tendency to overachieve. It often enriches text unnecessarily, ignoring requests for brevity and clarity where they're most needed.

LLaMA2 showed middling results, hesitating particularly in the *Number of Paragraphs* test with a -9.50 score. It was somewhat good at structuring content clearly, evidenced by a 0.35 in the *Clear Text Division test*, but its total score of -0.11 placed it among the bottom LLMs.

With inconsistent results, PaLM2 struggled notably with the *Simple Words* test but demonstrated some competency in sentence structuring and coherent content creation. However, its overall performance was dull, ending in a slightly better score than its cousin Bard at -0.32, reflecting its bad characteristics but counterintuitively benefitting from a blander approach.

The evaluation of the various LLMs reveals a clear disparity in performance and suitability for specialized tasks. GPT-4 distinguishes itself with its exceptional adaptability and strict adherence to rules, making it the superior choice for tasks requiring precision, such as those designed for individuals with CD. GPT-3.5 shows promise in word choice and conciseness, while LLaMA2 provides a somewhat balanced, albeit unspectacular, performance. On the other hand, Bard and PaLM2 fall significantly short of expectations. They not only show limited proficiency but also exhibit notable deficiencies in understanding and responding accurately to nuanced prompts, making them less reliable and ineffective for critical applications. Regarding tasks demanding strict adherence to prompt-based regulations, GPT-4 unequivocally stands out as the best option, overshadowing the underwhelming and inconsistent outputs of LLaMA2, PaLM2, and Bard.

4 CONCLUSIONS

We sought to explore the extent to which generative AI models could facilitate accessible interactions with individuals with cognitive disorders for the development of new generation educational tools. Following literature recommendations, we developed a set of guidelines to instruct large language models (LLMs) in creating accessible conversations. Subsequently, we conducted tests across five different topics to evaluate various LLMs' responsiveness to positive and negative instructions.

Our findings indicate that the majority of leading LLMs in the market exhibit sensitivity to instructions for accessible communication with people with cognitive disorders. Notably, GPT-based models, specifically GPT-3.5 and GPT-4, demonstrated the highest receptiveness during testing. This discovery marks a significant step towards the development of a new generation of educational tools that can effectively support users in meaningful conversations.

It is essential to recognize, however, that these findings are provisional. The capabilities of these models are currently undergoing exponential growth, continually integrating new features and improving overall performance. As such, we stress the importance of focusing on the outlined methodology and the discussion of results, prioritizing these aspects over the numerical outcomes. Decisions regarding model adoption should be reconsidered as the field evolves.

Moreover, our testing exclusively considered LLMs in their "as-they-are" state, without utilizing any fine-tuning techniques to enhance textual production. In future endeavors, we aim to construct a corpus of accessible texts and evaluate fine-tuned models using the same methodology, further advancing our understanding of generative AI's potential in facilitating accessible communication for individuals with cognitive disorders.

ACKNOWLEDGEMENTS

The authors are grateful to those who participated in our study at the "Fraternità e Amicizia" care center. This research is partially supported by the European Union – NextGenerationEU Program – National Recovery and Resilience Plan (NRRP), Mission 4, Component 2 Investment 1.4 – PNNR Project MUSA

– Multilayered Urban Sustainability Action – Spoke 6 (Innovation for Sustainable and Inclusive Societies) – Task 2.2 (Empowerment of persons with disabilities).

5 REFERENCES

[1] A. E. A. Lily, A. F. Ismail, F. M. Abunaser, F. Al-Lami e A. K. A. Abdullatif, "ChatGPT and the rise of semi-humans," May 20th, 2023. Retrieved from https://doi.org/10.1057/s41599-023-02154-3.

[2] K. Ciampa, Z. M. Wolfe e B. Bronstein, "ChatGPT in education: Transforming digital literacy practices," *Journal of Adolescent & Adult Literacy,* vol. 67, pp. 186-195, 2023.

[3] X. Zhai, "ChatGPT User Experience: Implications for Education," Dec 27th, 2022. Retrieved from https://ssrn.com/abstract=4312418.

[4] A. Zirar, "Exploring the impact of language models, such as ChatGPT, on student learning and assessment," *Review of Education,* 2023.

[5] F. Mosaiyebzadeh, S. Pouriyeh, R. Parizi, N. Dehbozorgi, M. Dorodchi e D. M. Batista, "Exploring the Role of ChatGPT in Education: Applications and Challenges.," *Association for Computing Machinery,* 2023.

[6] OpenAI, "GPT-4 Technical Report," 2023. Retrieved from https://arxiv.org/abs/2303.08774.

[7] A. Verma, S. Boland e K. Miesenberger, "Bridging the digital divide for persons with intellectual disabilities: assessing the role of chatgpt in enabling access, evaluation, integration, management, and creation of digital content," *16th annual International Conference of Education, Research and Innovation,* pp. 3767-3776, 2023.

[8] M. Melo, *ChatGPT as an Assistive Technology,* 2023. Retrieved from: https://www.insidehighered.com/views/2023/03/01/chatgpt-can-help-students-and-faculty-adhd-opinion.

[9] V. Gadiraju, S. Kane, S. Dev, A. Taylor, D. Wang, E. Denton e R. Brewer, " "I Wouldn't Say Offensive but...": Disability-Centered Perspectives on Large Language Models," *Association for Computing Machinery,* vol. 103, p. 205–216, 2023.

[10] GoogleAI, *An overview of Bard: an early experiment with generative AI,* 2023. Retrieved from https://ai.google/static/documents/google-about-bard.pdf.

[11] R. Anil et al., "PaLM 2 Technical Report", 2023. Retrieved from https://arxiv.org/abs/2305.10403.

[12] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample, "LLaMA: Open and Efficient Foundation Language Models," 2023. Retrieved from https://arxiv.org/abs/2302.13971.

[13] J. Qadir, "Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education," *Example,* 2023.

[14] K. Groom, "7 principles for engaging with neurodivergent people" Dec 10th, 2021. Retrieved from https://www.thedrum.com/opinion/2021/12/10/7-principles-engaging-with-neurodivergent-people.

[15] Neurodadversity, "Inclusive Language Guide – Neurodiversity Edition" Aug 8th, 2022. Retrieved from https://neurodiversitymatters.com/inclusive-language-guide-neurodiversity-edition/.

[16] P. Norvig e Google, *Natural Language Corpus Data: Beautiful Data*, 2009. Retrieved from https://norvig.com/ngrams/.

[17] L. S. Lo, *The Art and Science of Prompt Engineering: A New Literacy in the Information Age* Jun 23rd, 2023. Retrieved from: https://www.tandfonline.com/doi/full/10.1080/10875301.2023.2227621.

[18] M. R. Morris, "AI and accessibility," *Communications of the AC,* vol. 63, n. 6, pp. 35-37, 2020.

[19] F. Catania, F. Garzotto e M. Spitale, "Conversational Agents in Therapeutic Interventions for Neurodevelopmental Disorders: A Survey," *Association for Computing Machinery,* vol. 55, n. 10, pp. 1-34, 2023.

[20] M. Houben, N. van As, N. Sawhney, D. Unbehaun e M. Lee, "Participatory Design for Whom? Designing Conversational User Interfaces for Sensitive Settings and Vulnerable Populations," *Association for Computing Machinery,* n. 60, pp. 1-4, 2023.

[21] L. Seeman-Horwitz, R. B. Montgomery, S. Lee e R. Ran, "Making Content Usable for People with Cognitive and Learning Disabilities," Apr 29th, 2021. Retrieved from https://www.w3.org/TR/2021/NOTE-coga-usable-20210429/.

[22] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer e Oleksandr, "ChatGPT for good? On opportunities and challenges of large language models for education,".

[23] S. S. Gill, M. Xu, P. Patros, H. Wu, R. Kaur, K. Kaur, S. Fuller, M. Singh, P. Arora, A. K. Parlikad, V. Stankovski, A. Abraham, S. K. Ghosh, H. Lutfiyya e S. Kanhere, "Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots," *Internet of Things and Cyber-Physical Systems,* vol. 4, pp. 19-23, 2024.

[24] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi e G. Neubig, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing."

[25] N. Rane, "Chatbot-Enhanced Teaching and Learning: Implementation Strategies, Challenges, and the Role of ChatGPT in Education," Jul 21st, 2023. Retrieved from https://ssrn.com/abstract=4603204.

[26] N. Rane, "Roles and Challenges of ChatGPT and Similar Generative Artificial Intelligence for Achieving the Sustainable Development Goals (SDGs)," Aug 4th, 2023. Retrieved from https://ssrn.com/abstract=4603244.

[27] L. Giray, "ChatGPT References Unveiled: Distinguishing the Reliable from the Fake," Internet Reference Services Quarterly, Retrieved from https://doi.org/10.1080/10875301.2023.2227621.

[28] J. Goulet, "Stop Asking Neurodivergent People to Change the Way They Communicate," Oct, 5th 2022. Retrieved from https://hbr.org/2022/10/stop-asking-neurodivergent-people-to-change-the-way-they-communicate.