Human-AI Heuristic Evaluation: Uncovering usability insights of an LLM Chatbot Interface for Personalized Learning in Autism

Yuri P. S. Zaidan, Eric F. Monteiro, Duncan Ruiz, Afonso Sales, Milene Silveira

[1]School of Technology – PUCRS
Avenida Ipiranga, 6681 – 90619-900 – Porto Alegre, RS – Brazil

{y.zaidan, f.eric}@edu.pucrs.br, {duncan.ruiz, afonso.sales, milene.silveira}@pucrs.br

Abstract. *The rise of Large Language Models (LLMs) and Generative AI (GenAI) offers new possibilities for personalized learning but also introduces usability challenges, especially in applications designed to customize the learning experience for individuals with autism spectrum disorder (ASD). To analyze whether an LLM-powered chatbot interface is effective for this specific use case, we propose a usability heuristics inspection carried out by both human experts and an AI agent to evaluate an LLM interface for ASD curriculum personalization. Preliminary results reveal human and AI-driven usability insights towards a frictionless experience for GenAI educational interventions.*

1. Introduction

With 2.4 million individuals diagnosed with autism spectrum disorder (ASD), the landscape of autism care in Brazil is undergoing significant changes [Instituto Brasileiro de Geografia e Estatística 2024]. The financial implications are substantial; in 2024, the Brazilian Association of Healthcare Providers noted that autism treatment costs surpassed cancer treatment expenditures for the first time [Fenacor 2025]. This underscores the increasing need for medical interventions that support cognitive, language, and behavioral development in this population.

ASD encompasses a wide range of learning profiles, with individuals exhibiting diverse strengths and challenges that often require highly personalized educational strategies [Pellicano et al. 2014]. Furthermore, foundational research in ASD has consistently demonstrated the benefits of visual supports for learners across such spectrum. These supports have shown considerable promise in making abstract concepts more accessible, reducing anxiety, and fostering engagement within educational settings for the ASD community [Tissot and Evans 2003].

LLMs and AI agents have emerged as powerful tools for generating and adapting educational content [Kasneci et al. 2023]. For those providing educational support to individuals with ASD, who often devote significant time crafting personalized visual aid resources [Kohli et al. 2022], LLM chatbot interfaces offer the potential to automate and customize these materials efficiently, allowing for more responsive and individualized support within learning interventions.

In view of the latest developments in the fields of LLMs and ASD education, the research on Human-Computer Interaction (HCI) offers essential frameworks for evaluating the usability of emerging technologies, particularly as they are applied to design

DOI: 10.5753/sbie.2025.12579

individualized educational content. As LLMs enable unprecedented levels of personalization in educational and therapeutic resources, ensuring these tools are intuitive and accessible becomes critical. Usability heuristics, such as those proposed by Jakob Nielsen [Nielsen 1994], provide valuable benchmarks for assessing whether LLM-based systems truly meet the diverse needs of educators and therapists seeking to personalize ASD curriculum.

Drawing on the literature of heuristic evaluation, this study aims to answer the following research question (RQ):

How do human and AI heuristic evaluations compare in identifying usability challenges of LLM chatbot interfaces for personalized learning in autism?

To address the question outlined above, this work first reviews literature relevant to this research agenda, then outlines the methodology adopted for the study, followed by a discussion of the emerging results obtained. Finally, it highlights the study's limitations and suggests directions for future research.

2. Background and Related Work

ASD can influence how individuals perceive, process, and respond to information, often resulting in distinct learning journeys that require careful consideration. Many learners with ASD experience difficulties with abstract reasoning, attention, and adapting to traditional instructional methods, which can hinder their academic and social development [Christensen and Zubler 2020]. These obstacles underscore the necessity for highly personalized interventions that can adapt to each learner's evolving needs and preferences [Happé and Frith 2006]. As a result, educators and therapists are increasingly seeking methods that allow for flexible and individualized instruction [Hume et al. 2021].

Research on ASD consistently emphasizes the importance of centering educational approaches around the individual needs and preferences of each apprentice with ASD [Carvalho et al. 2024]. Rather than relying on one-size-fits-all methods, researchers and clinicians advocate for flexible strategies that adapt to the learner's strengths, challenges, and interests. In this context, assistive technologies — including visual supports and digital tools — have gained recognition for their ability to personalize instruction and promote greater engagement, making learning more accessible and meaningful for individuals on the spectrum [Cunha and Carvalho 2024].

In this sense, Artificial Intelligence (AI) serves as the foundation for a range of advanced technologies, including Generative AI and LLMs [Brown et al. 2020]. Generative AI, powered by LLM's chatbots, can analyze vast amounts of data and generate tailored content, making it possible to design educational interventions that are responsive to individual learning profiles [Ng and Fung 2024]. By leveraging these technologies, educators and therapists can create resources that adapt in real time to the needs, preferences, and progress of each learner, engaging throughout their learning journey. Building on this foundation, emerging studies have shown how LLMs are being used to generate highly customized curricula for learners with ASD [Carik et al. 2025, Papadopoulos 2024, Shi et al. 2024].

Accordingly, HCI has become increasingly critical with the proliferation of LLM tools, as these interfaces mediate how users interact with complex AI systems

[Quéré et al. 2025]. HCI principles guide the design and evaluation of LLM interfaces to ensure they are intuitive, efficient, and effective for a diverse range of users [Morris 2025]. As LLMs become more integrated into various professional domains, HCI offers a framework for optimizing user experiences and maximizing the potential benefits of these technologies.

Expanding on this perspective, Nielsen's ten usability heuristics (UH) provide a structured approach to assess the usability of LLM interfaces, offering a set of established guidelines for evaluating design elements and identifying potential usability issues [Nielsen 1994]. These heuristics, often based on principles of cognitive psychology and user-centered design, enable researchers and practitioners to systematically examine how well an interface supports user goals and minimizes frustration [Aubin Le Quéré et al. 2024]. Ultimately, by applying usability heuristics, researchers can potentially access valuable insights into the strengths and weaknesses of LLM interfaces, leading to iterative improvements and enhanced user satisfaction.

Analyzing the usability of LLM interfaces is particularly important for professionals who plan to use these resources for personalizing learning interventions. Educators and therapists need LLM tools that are not only powerful but also easy to learn and use effectively. A well-designed LLM interface can streamline the process of creating and adapting personalized learning materials, allowing professionals to focus on the unique needs of their learners rather than struggling with complex technology [Hao et al. 2022].

Addressing this challenge, AI agents are autonomous systems designed to perceive their environment, make decisions, and take actions to achieve specific goals, often adapting their behavior based on feedback [Wang et al. 2024]. Unlike traditional software, which follows fixed rules and requires explicit instructions, AI agents can learn from data, reason about complex scenarios, and operate with a degree of independence that allows for dynamic problem-solving [Xi et al. 2023]. This flexibility marks a significant departure from static, rule-based approaches.

Moreover, AI agents can be harnessed to enhance the evaluation of usability heuristics in digital tools. By drawing on the expertise embedded in large language models, these agents can systematically analyze interfaces, identify potential usability issues, and provide detailed feedback based on established heuristic principles while simulating the perspectives of specialists — such as ASD therapists or educators — when assessing the usability of learning technologies. This approach allows AI agents to not only streamline the evaluation process but also ensure that feedback is grounded in domain-specific knowledge, making usability assessments more relevant and actionable.

3. Method

To address the central question of this study, we propose a dual heuristic evaluation of an LLM chatbot interface, as illustrated in Figure 1. Two human experts with more than 6 years of experience developing interactive web systems and an AI agent independently assessed the interface, allowing for a comparison of the usability friction points identified by each when creating custom educational resources through an LLM tool.

Nielsen's usability heuristics were evaluated against each of the three stages of the navigational workflow: *main interface, prompt entry process, and response interface* as

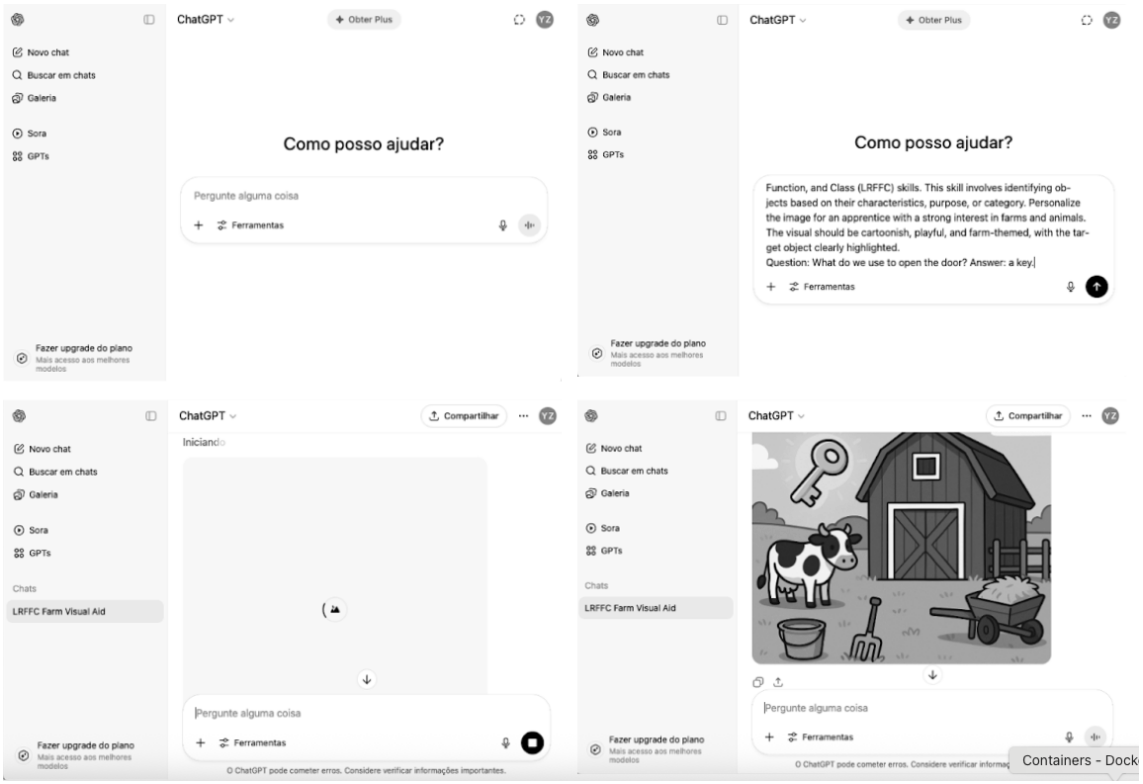

Figure 1. Different states of the ChatGPT LLM chatbot interface inspected by humans experts and the AI agent

shown in Figure 2. This group of heuristics was chosen because it was designed to be simple, intuitive, and applicable across a wide range of interactive systems. It provides a framework for evaluators to assess the usability of an interface from the user's perspective, spotting friction points that can negatively impact the user experience [Nielsen 1994].

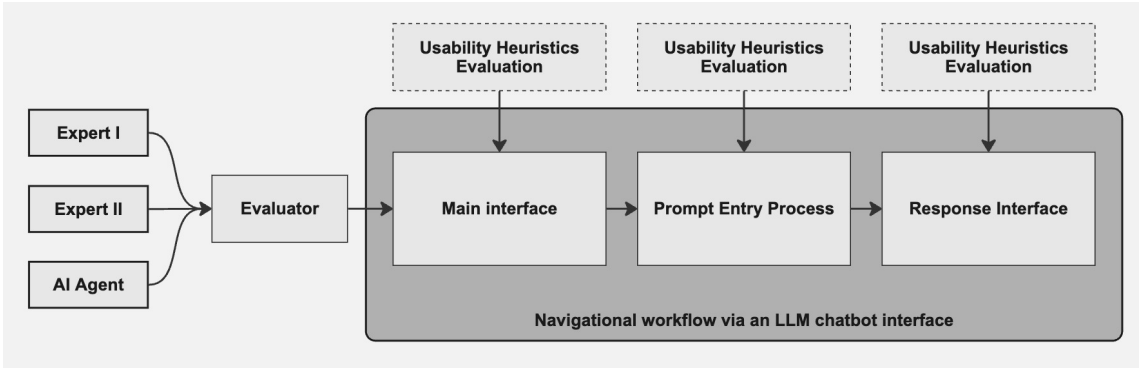

Figure 2. Research diagram

Given the widespread dissemination of OpenAI's LLM Chat Generative Pre-trained Transformer (ChatGPT), this study chose to evaluate its chatbot interface. The latest GPT model available through the free tier at the time of this research was GPT-4o [OpenAI 2024a]. Therefore, both the AI agent and the human experts used the model version GPT-4o during this usability evaluation.

A standardized text prompt was engineered based on the guidelines provided by

OpenAI [OpenAI 2024b] describing the request to the LLM chatbot. In this case, the LLM chatbot was asked to create a visual aid resource as shown in Table 1.

Table 1. Text used by experts and AI agent to prompt LLM

Text Prompt
Context: You are an expert in autism spectrum disorder (ASD). Your role is to personalize visual educational resources that will support learning interventions targeted at learners with ASD. **Goal:** Design a visual aid to assess Listener Responding by Feature, Function, and Class (LRFFC) skills. This skill involves identifying objects based on their characteristics, purpose, or category. Personalize the image for an apprentice with a strong interest in farms and animals. The visual should be cartoonish, playful, and farm-themed, with the target object clearly highlighted. **Question:** *What do we use to open the door? Answer: a key.*

Abacus.AI's general-purpose agent DeepAgent is an AI agent based on a multi-LLM architecture that leverages a variety of state-of-the-art models [Abacus.AI 2024]. This agent was chosen to perform the usability heuristics evaluation of the ChatGPT chatbot interface using the commercially available version of DeepAgent on May 31, 2025. The AI agent was prompted to perform the heuristic evaluation based on the instructions shown by Table 2.

Table 2. Text used to prompt the AI Agent to perform a usability heuristics evaluation of the ChatGPT chatbot interface

Text Prompt
Context: You're an expert in user experience design and you are tasked to access ChatGPT chatbot interface as an autism spectrum disorder (ASD) therapist who will prompt the LLM chatbot interface to create a custom visual aid resource for an educational program based on a profile of an apprentice with ASD. **Goal:** Evaluate ChatGPT's usability using Nielsen's 10 heuristics as you move through three key stages of the interaction flow: a) Main interface b) Prompt-entry process c) Response interface Begin by navigating to the ChatGPT homepage. You will first be prompted to sign in with a Microsoft account; only after logging in will the "Create Image" option becomes available. (Use the designated Microsoft account credentials provided to you.) **Prompt:** *Text prompt shown by Table 1.*

4. Results

Preliminary findings for each usability heuristic were recorded in an evaluation template table. Both human experts and the AI Agent then analyzed each stage of the navigational workflow to identify potential obstacles that could hinder educators and therapists from experiencing a frictionless interaction with the LLM chatbot interface.

To illustrate the results, Table 3 presents the emerging findings related to the "Visibility of System Status" heuristic, examined across the main interface, prompt entry process, and response interface of the chatbot. At each stage of navigation, both human experts and the AI agent independently evaluated the interface and documented their observations[1].

Table 3. Results of heuristic evaluation: Visibility of system status

Usability heuristic: Visibility of system status			
Workflow stage	**Expert I**	**Expert II**	**AI Agent**
Main Interface	The system does not display the overall system status.	The model change indicator (*e.g.*, GPT-4o or GPT-3.5) is easy to miss and lacks an explanation of what the model version change implies.	ChatGPT's usability is good due to clear status indicators, but adding estimated processing times would further help therapists plan workflows.
Prompt Entry Process	While the overall status is visible, adding a clear call-to-action button to initiate the chat would greatly benefit users who are less tech-savvy.	The "Run deep research" button does not clarify what is a deep research.	ChatGPT's text input area is visually clear but could benefit from word count and prompt complexity indicators to help therapists optimize content generation.
Response Interface	The system lacks an estimated response time, particularly when handling multiple user requests, which can leave users uncertain about processing duration.	Some responses include technical language (*e.g.*, "token limit exceeded") without clarification for average users.	Therapists could benefit from progressive text generation to preview content quality early in the generation process, enabling them to interrupt the process.

The comparative analysis of the usability heuristics inspection for the ChatGPT chatbot interface, as presented in Table 3, reveals both convergences and divergences between the evaluations conducted by human experts and the AI agent. Both human experts identified gaps in the visibility of system status, such as the absence of clear overall status indicators and estimated response times, which can leave users uncertain about the system's current state and processing duration.

[1]The complete table with detailed analyses is available via the external link `https://brpucrs-my.sharepoint.com/:x:/g/personal/y_zaidan_edu_pucrs_br/ERR6y0WoWetFvnPUz9Zcc0YBXb5bAIYBBEMYFuKKJVuYsw?e=66fCLa`

Expert I emphasized the need for clearer call-to-action elements and highlighted the potential confusion for less tech-savvy users, while Expert II noted the lack of explanatory context for model changes and the use of technical jargon in system messages. The AI agent's evaluation, although generally more favorable about the existing status indicators, also acknowledged the benefit of incorporating estimated processing times. Additionally, it recommended features such as word count and prompt complexity indicators to better support user workflows when generating the expected custom visual aid resources (see Figure 3).

Notably, the AI agent proposed progressive text generation as a means to improve user control during response delivery, aligning with the experts' concerns about transparency and user feedback. Overall, the AI agent demonstrated the ability to generate coherent and relevant usability insights, echoing many of the human experts' observations while also introducing additional, actionable suggestions. This suggests that AI-driven heuristic evaluations can complement human expertise, particularly in identifying opportunities for workflow optimization and user empowerment within conversational interfaces.

Figure 3 displays the visual aids generated by the LLM, customized to align with the specific needs and interests of learners with ASD. These resources were created in response to prompts provided by human experts and the AI agent as part of their heuristic evaluation carried out in this study.

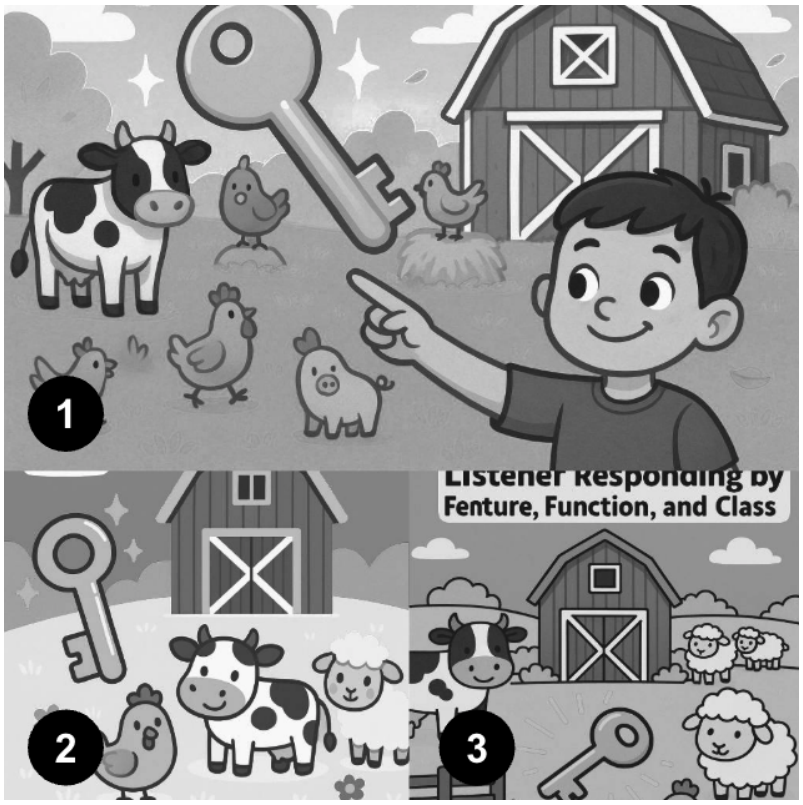

Figure 3. Visual aid resources created by human experts (Images 1 and 2) and AI agent (Image 3)

5. Final Considerations

This study was motivated by the growing need for effective and accessible personalized learning tools for individuals with autism spectrum disorder (ASD), particularly in light of the increasing adoption of Large Language Models (LLMs) and generative AI in educational contexts. Recognizing the unique learning profiles and challenges faced by individuals with ASD, the research focused on evaluating the usability of an LLM-powered chatbot interface designed to support the creation of customized educational resources. To address the central research question — how human and AI heuristic evaluations compare in identifying usability challenges of LLM chatbot interfaces for personalized learning in autism — the study employed a dual heuristic inspection approach.

Two experienced human experts and an AI agent independently assessed the ChatGPT interface using Nielsen's usability heuristics across key stages of the user workflow. The findings indicate that the AI agent was able to generate coherent and relevant usability insights, often aligning with the observations of human experts while also contributing additional actionable suggestions. These results suggest that AI-driven heuristic evaluations can effectively complement human expertise, offering a promising avenue for enhancing the usability of LLM interfaces in personalized learning for ASD. This comparative analysis is a core contribution of the research, and future studies should delve deeper into identifying where AI excels (*e.g.*, identifying subtle patterns in code) and where human expertise is indispensable (*e.g.*, understanding the complex social and cognitive needs of individuals with ASD.

To advance this research, the usability evaluation must be expanded to include a more diverse range of perspectives. The evaluation panel should be expanded to include Autism therapists and educators. It is essential to define a comparable scope of analysis across all examiners — human experts, therapists, and AI agents — to ensure evaluations measure the same aspects of usability, allowing for a direct comparison of insights.

Future research should aim to quantify the friction identified by Nielsen's heuristics from the therapist's perspective when personalizing learning resources. This can be achieved by developing a scoring system or qualitative scale to measure the difficulty, cognitive load, or frustration experienced by therapists. It is also critical to highlight findings that emerge specifically from the comparison between AI-generated and human-generated assessments.

As for the diversification of LLMs and AI agents, future research should also include a wider range of LLMs. Evaluating various models (*e.g.*, different versions of GPT, or open-source models like Llama) will provide a more robust understanding of the consistency and reliability of AI-driven heuristic evaluations. This step will help determine if the effectiveness of this method is dependent on the specific AI model used or if it represents a generally applicable approach.

Acknowledgment

This study was partially supported by the Ministry of Science, Technology, and Innovations from Brazil, with resources from Law No. 8.248, dated October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex, and by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) and PROEX.

References

Abacus.AI (2024). DeepAgent. Available in: https://deepagent.abacus.ai/. Access in May, 2025.

Aubin Le Quéré, M., Schroeder, H., Randazzo, C., Gao, J., Epstein, Z., Perrault, S. T., Mimno, D., Barkhuus, L., and Li, H. (2024). LLMs as Research Tools: Applications and Evaluations in HCI Data Work. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Carik, B., Ping, K., Ding, X., and Rho, E. H. (2025). Exploring Large Language Models Through a Neurodivergent Lens: Use, Challenges, Community-Driven Workarounds, and Concerns. *Proc. ACM Hum.-Comput. Interact.*, 9(1).

Carvalho, E., Alves, F., Rodrigues, I., Souza, T., and Moreira, D. (2024). Autismo e Tecnologias Assistivas: uma Revisão Sistemática dos Anais do Congresso Brasileiro de Informática na Educação. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, pages 1084–1098, Porto Alegre, RS, Brasil. SBC.

Christensen, D. and Zubler, J. (2020). CE: From the CDC: Understanding Autism Spectrum Disorder. *American Journal of Nursing*, 120(10):30–37.

Cunha, M. and Carvalho, L. (2024). ABC Autismo Frutas: Um aplicativo para crianças com autismo construído com base nas premissas do Design Centrado no Usuário e do Ensino Estruturado. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, pages 937–950, Porto Alegre, RS, Brasil. SBC.

Fenacor (2025). Fenacor. Available in: https://www.fenacor.org.br/noticias/autismo-supera-cancer-em-custos-de-planos-de. Access in May, 2025.

Hao, Y., Song, H., Dong, L., Huang, S., Chi, Z., Wang, W., Ma, S., and Wei, F. (2022). Language Models are General-Purpose Interfaces. arXiv preprint arXiv:2206.06336.

Happé, F. and Frith, U. (2006). The Weak Coherence Account: Detail-focused Cognitive Style in Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, 36(1):5–25.

Hume, K., Steinbrenner, J. R., Odom, S. L., Morin, K. L., Nowell, S. W., Tomaszewski, B., Szendrey, S., McIntyre, N. S., Yücesoy-Özkan, S., and Savage, M. N. (2021). Evidence-Based Practices for Children, Youth, and Young Adults with Autism: Third Generation Review. *Journal of Autism and Developmental Disorders*, 51(11):4013–4032.

Instituto Brasileiro de Geografia e Estatística (2024). Censo 2022: Indicadores - Educação. Available in: https://censo2022.ibge.gov.br/panorama/indicadores.html?localidade=BR&tema=9. Access in May, 2025.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., and Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Kohli, M., Kar, A. K., Bangalore, A., and Ap, P. (2022). Machine learning-based ABA treatment recommendation and personalization for autism spectrum disorder: an exploratory study. *Brain Informatics*, 9(1):16.

Morris, M. R. (2025). HCI for AGI. *Interactions*, 32(2):26–32.

Ng, C. and Fung, Y. (2024). Educational Personalized Learning Path Planning with Large Language Models. arXiv preprint arXiv:2407.11773.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, page 152–158, New York, NY, USA. Association for Computing Machinery.

OpenAI (2024a). Hello GPT-4o. Available in: `https://openai.com/index/hello-gpt-4o/`. Access in May, 2025.

OpenAI (2024b). Prompt Engineering — OpenAI Platform Documentation. Available in: `https://platform.openai.com/docs/guides/text?api-mode=chat#prompt-engineering`. Access in May, 2025.

Papadopoulos, C. (2024). Large language models for autistic and neurodivergent individuals: Concerns, benefits and the path forward. *Neurodiversity*, 2:27546330241301938.

Pellicano, E., Dinsmore, A., and Charman, T. (2014). What should autism research focus upon? Community views and priorities from the United Kingdom. *Autism*, 18:756–770.

Quéré, M. A. L., Schroeder, H., Randazzo, C., and Gao, J. (2025). The State of Large Language Models in HCI Research: Workshop Report. *Interactions*, 32(1):8–9.

Shi, Z., Landrum, E., O'Connell, A., Kian, M., Pinto-Alva, L., Shrestha, K., Zhu, X., and Matarić, M. J. (2024). How Can Large Language Models Enable Better Socially Assistive Human-Robot Interaction: A Brief Survey. *Proceedings of the AAAI Symposium Series*, 3(1):401–404.

Tissot, C. and Evans, R. (2003). Visual Teaching Strategies for Children with Autism. *Early Child Development and Care*, 173(4):425–433.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., and Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6).

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., and Gui, T. (2023). The Rise and Potential of Large Language Model Based Agents: A Survey. *arXiv preprint arXiv:2309.07864*.