frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 15 October 2024
DOI 10.3389/frai.2024.1460364

OPEN ACCESS

EDITED BY
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain

REVIEWED BY
Thomas Mandl,
University of Hildesheim, Germany
Zhi Liu,
Central China Normal University, China

*CORRESPONDENCE
Noah L. Schroeder
✉ schroedern@ufl.edu

RECEIVED 05 July 2024
ACCEPTED 30 September 2024
PUBLISHED 15 October 2024

CITATION
Mannekote A, Davies A, Pinto JD, Zhang S, Olds D, Schroeder NL, Lehman B, Zapata-Rivera D and Zhai C (2024) Large language models for whole-learner support: opportunities and challenges.
Front. Artif. Intell. 7:1460364.
doi: 10.3389/frai.2024.1460364

Large language models for whole-learner support: opportunities and challenges

Amogh Mannekote[1], Adam Davies[2], Juan D. Pinto[3], Shan Zhang[4], Daniel Olds[5], Noah L. Schroeder[1]*, Blair Lehman[6], Diego Zapata-Rivera[6] and ChengXiang Zhai[2]

[1]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, United States, [2]Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL, United States, [3]Department of Curriculum and Instruction, University of Illinois Urbana-Champaign, Champaign, IL, United States, [4]School of Teaching and Learning, University of Florida, Gainesville, FL, United States, [5]Department of Computer Science, University of Oregon, Eugene, OR, United States, [6]Educational Testing Service, Princeton, NJ, United States

In recent years, large language models (LLMs) have seen rapid advancement and adoption, and are increasingly being used in educational contexts. In this perspective article, we explore the open challenge of leveraging LLMs to create personalized learning environments that support the "whole learner" by modeling and adapting to both cognitive and non-cognitive characteristics. We identify three key challenges toward this vision: (1) improving the interpretability of LLMs' representations of whole learners, (2) implementing adaptive technologies that can leverage such representations to provide tailored pedagogical support, and (3) authoring and evaluating LLM-based educational agents. For interpretability, we discuss approaches for explaining LLM behaviors in terms of their internal representations of learners; for adaptation, we examine how LLMs can be used to provide context-aware feedback and scaffold non-cognitive skills through natural language interactions; and for authoring, we highlight the opportunities and challenges involved in using natural language instructions to specify behaviors of educational agents. Addressing these challenges will enable personalized AI tutors that can enhance learning by accounting for each student's unique background, abilities, motivations, and socioemotional needs.

KEYWORDS

large language model (LLM), AI and education, non-cognitive aspects of learning, interpretability, pedagogical support of students, educational authoring tool

1 Introduction

In recent years, generative artificial intelligence (GenAI)—and more specifically, LLMs—have exploded into global public awareness (Barreto et al., 2023). ChatGPT, for example, is available in 188 countries with over 180 million users (as of August 2023)[1]. Such rapid adoption and ongoing development continues to disrupt many industries and areas of study, particularly as each new generation of LLMs offers new capabilities (e.g., memory, multimodality, longer input context sizes). LLMs have made their impact in the world of education as well—for instance, one notable example is Khanmigo[2], an LLM-powered AI

1 https://clickup.com/blog/chatgpt-statistics/#6-chatgpt-users-and-usage-

2 https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/

tutor that provides personalized support to students and assists teachers with developing instructional materials. This type of personalized support highlights the great potential for LLMs in educational contexts (Pardos and Bhandari, 2023).

We argue that such personalized support systems can and should be further expanded to provide "whole learner" support, moving beyond the paradigm of understanding and supporting only students' *academic proficiency* to also address *social*, *affective*, *motivational*, *cultural*, and *linguistic* characteristics that are known to impact learning (Bernacki et al., 2021; Mercado, 2018; Walkington and Bernacki, 2018; Lehman et al., 2024). This work focuses on "non-cognitive skills," describing aspects of a learner beyond subject knowledge and proficiency, such as resilience, persistence, empathy, motivation, self-regulation, and a growth mindset (Kautz et al., 2014).

Personalized learning environments (PLEs) leverage learner models, which are structured representations of students, to guide personalized support (Abyaa et al., 2019; Ismail et al., 2023). However, existing learner models cannot support the "whole learner," as they typically limit themselves to modeling knowledge acquisition (e.g., level of mastery over a concept), and at best, one additional characteristic (e.g., prior knowledge) or behavior (e.g., engagement; Ismail et al., 2023)[3]. The complexity of whole-learner modeling stems from the fact that it is not enough to simply model each characteristic and behavior independently—instead, these factors must be considered holistically to understand and support the whole learner. While the complex interaction of these factors presents a significant challenge for existing PLEs, we know that such holistic support is possible to provide in practice given that human teachers successfully combine these elements to support their students every day.

By pairing whole-learner modeling with GenAI, LLMs present an opportunity to bridge the long-standing gap between the quality and range of learner support offered by present-day computational systems and that offered by expert human tutors (Lepper et al., 1993; Cade et al., 2008; D'Mello et al., 2010). However, rapid innovation in the field of LLMs has raised questions about their appropriate use in PLEs. We explore some of the challenges and opportunities that exist around the vision of using LLMs to build whole-learner models and eventually create adaptive learning systems. We first explore the challenges and potential of LLMs in doing so (Section 2) and then identify several promising research directions to address these challenges (Section 3).

2 Challenges and the state of the art

We identify three key areas where the community needs to progress to achieve the larger vision of whole-learner modeling:

- **Interpretable representations of learners:** It is necessary to represent a learner explicitly and faithfully, including both *cognitive* and *non-cognitive* aspects. Although deep learning methods have traditionally been viewed as "black-box" approaches with opaque internal mechanisms, recent advances in interpretability and explainability research are working to address this challenge, and are well-positioned for applications in the context of whole learner modeling and support.
- **Adaptive technologies to support whole learners:** Given an interpretable learner representation, it should be used to tailor the delivery of pedagogical content and support to suit a learner's characteristics, including both cognitive and non-cognitive states. By leveraging data on learner behavior, preferences, and characteristics, and dynamically adjusting instructional strategies to address individual needs, adaptive systems can provide personalized learning pathways that evolve with learners' cognitive and non-cognitive skill development.
- **Authoring and evaluating agents:** In the context of PLEs and pedagogical agents (PAs), the term "authoring" refers to the manual process of specifying behaviors (or "policies") of the agent. For instance, in an intelligent tutoring system (ITS), classroom instructors can *author* common misconceptions based on their teaching experience. Authors can come from varying backgrounds (e.g., researcher, educator, and developer), so a key challenge in designing authoring tools is balancing accessibility and minimizing cognitive load. Finally, authoring is tightly-coupled with the issue of evaluation, a critical step in smoothly deploying these systems to real learners.

2.1 Interpretable representations of learners

Two key concerns in deploying LLMs to potentially sensitive application contexts such as education are *interpretability* (what a "black box" model is doing and representing internally) and *explainability* (why the model outputs A instead of B, given input C). Without reliable *interpretation*, we do not know what information the models use to make decisions or generate responses. Unlike trained educational professionals, automated models cannot be trusted to reliably take students' prior knowledge or emotional state into account to provide relevant and compassionate guidance, nor can we be certain that they will not use sensitive demographic information inappropriately. On the other hand, when we cannot reliably *explain* LLM behaviors, we cannot ensure that desired behaviors in one context will generalize to others (e.g., whether attentiveness to the emotional needs of students with high socioeconomic status will translate to less advantaged students).

Integrating interpretable learner models with LLMs is a promising approach to develop PLEs, providing the benefits of GenAI while maintaining a high level of interpretability. Such a hybrid approach need not be overly complex; for instance, one may begin by training a traditional learner model and passing its inferences to the LLM as an additional component of input prompts. However, it is crucial to ensure that (1) LLMs actually consider learner model's output, and that (2) they use this information in a way that is faithful to the learner model and consistent with educational best practices – otherwise, the approach

3 However, models considering more learner characteristics have been proposed (Shute and Zapata-Rivera, 2008; Zapata-Rivera and Greer, 2004).

will not benefit educational stakeholders like teachers and students (Pinto et al., 2023).

LLMs are also well-suited for advancing open learner models (OLMs) due to their natural language dialogue capabilities. OLMs enable learners to view and interact with the model's representation of their knowledge, promoting reflection and self-regulated learning. This transparency and interactivity can enhance traditional OLMs, allowing learners to modify their learning paths more freely. However, the use of LLMs in OLMs also raises concerns about how LLMs use learner information and relate their actions to educational best practices. Indeed, the integration of LLMs with OLMs has the potential to revolutionize educational technology by making learning processes more adaptive and personalized (Conati et al., 2018; Kay et al., 2020; Zapata-Rivera and Arslan, 2021; Bull, 2020), but implementation must be guided by strong ethical and pedagogical standards.

Despite important recent advances in understanding the inner workings of LLMs (e.g., Elhage et al., 2021; Olsson et al., 2022; Conmy et al., 2023; Templeton et al., 2024), reliably explaining model behavior to relevant stakeholders remains a significant challenge. This inability to interpret LLM representations and explain model behaviors leads to a lack of trust (Shin, 2021; Liao and Vaughan, 2023), which can inhibit these models' deployment to educational contexts where they have potential for transformational impact. By contrast, many traditional learner models are designed with interpretability as an inherent feature, such as Bayesian Knowledge Tracing (Baker et al., 2008) and Item Response Theory (Yen and Fitzpatrick, 2006). There are even efforts underway to develop intrinsically interpretable neural-network-based learner models (Pinto et al., 2023; Lu et al., 2020; Swamy et al., 2024). We discuss how such approaches can address the challenges of interpretable LLMs for education in **3.1**.

2.2 Whole learner support through adaptive technologies

The next challenge in deploying LLMs for education is *adaptivity*, which involves assessing various learner characteristics and tailoring the learning experience to individual needs to improve learning outcomes (Plass and Pawar, 2020). Through the natural language capabilities of LLMs, adaptive technologies for whole learner support can offer nuanced support for developing both cognitive and non-cognitive skills in diverse learners. For instance, Arroyo et al. (2014) demonstrated that intelligent adaptive tutors effectively address students' unique needs and emotions, enhancing engagement and affect, while Liu et al. (2024) found that conversational agents offering emotional scaffolding improved students' emotional experiences.

Such findings highlight the importance of design principles focused on non-cognitive learner characteristics, such as fostering a growth mindset through praising learners' efforts (Liu et al., 2024), attributing struggles to external factors (Calvo and D'Mello, 2011), utilizing an anthropomorphic language style, and employing proactive inquiry (McQuiggan et al., 2008; Sabourin et al., 2011) to guide learners to self-report their emotional states. For instance, a review found that empathetic agent feedback, including affective feedback and confidence- and motivation-enhancing dialogue, positively influences students' attitudes (Liu et al., 2024; Ortega-Ochoa et al., 2024). Similarly, another study demonstrated how conversational agents can support children's social-emotional learning by teaching self-talk. These lines of research also emphasize the importance of designing conversational dialogue based on an evidence-based framework (Fu et al., 2023). Building on this foundation, recent AI advancements have facilitated the development of natural language dialogue systems to scaffold non-cognitive skills (Acosta et al., 2015; Anghel and Balart, 2017; Cinque et al., 2021).

2.3 Authoring and evaluating agents

LLMs are also transforming the landscape for authoring educational agents such as PAs, intelligent tutors (Sottilare et al., 2015), and even simulated learners (Käser and Alexandron, 2023). Before the widespread adoption of modern LLMs, agent authoring was bottlenecked by supervised and reinforcement learning methods that required machine learning expertise (Mannekote et al., 2023; Liu and Chilton, 2022), lots of data, labor-intensive manual annotation, or some combination of these factors. In contrast, the recent development of instruction-tuned LLMs (Wang et al., 2023) enables educational experts to define agent behaviors using natural language instructions in "zero-shot" or "few-shot" setups (i.e., using no annotated examples or only a few, respectively). In addition to reducing the training and expertise needed for authoring the dialogue system, LLMs also open up new avenues of agent behavior—for instance, where classical ITSs predominantly focused on supporting the cognitive aspects of learning (e.g., subject proficiency) (Sottilare et al., 2015), authors can now leverage LLMs capabilities such as their abilities to emulate human-like decision-making (Miliv cka et al., 2024) and perform high-level planning (Kambhampati et al., 2024) to equip them with the ability to support non-cognitive aspects of learning as well.

However, LLMs are not (yet) a "turn-key" solution to agent authoring, as several key challenges remain. Authoring LLM-based agents requires effectively navigating an unbounded space of possible prompts, which may be difficult to do without prompt engineering expertise (Oppenlaender et al., 2023; Zamfirescu-Pereira et al., 2023; Mannekote et al., 2023). Moreover, it has been shown that LLM outputs are highly sensitive to minor prompt variations, often leading to inconsistent (Lu et al., 2022; Liu and Chilton, 2022; Loya et al., 2023; Mohammadi, 2024) and confounding (Gui and Toubia, 2023) results. Finally, when authoring complex agent behaviors, the issue of evaluating the *faithfulness* of an agent's behavior to the authors' intended expectations becomes pertinent (Koedinger et al., 2015; Weitekamp et al., 2023). In fact, within the context of AI models like LLMs, this issue can be considered to be a specific instance of the *alignment problem* (Yudkowsky, 2016).

3 Ways forward

For each of the three challenge areas delineated in Section 2, we outline a broad roadmap for future advancements. Specifically, we

identify promising directions that the field is likely to pursue in the medium to long term.

3.1 Interpretable representations of learners

We face two primary challenges in enhancing the interpretability of LLMs. First, rather than merely adding more information to the prompt and hoping that the model will use it appropriately, we need a direct method to explain why a model generated a particular output. This involves determining whether the output was produced due to explicit learner information that has been added to prompts or implicit learner information that LLMs have inferred from learner behavior. Second, we must predict whether its behavior will remain consistent when applied in different contexts, such as in learning environments for which they have not already been tested. Although neural networks like LLMs are usually seen as "black boxes" whose internal representations and mechanisms are treated as unknowable beyond the outputs they produce, recent work in deep learning interpretability has made substantial strides in addressing this challenge. For instance, current interpretability methods can detect what latent representations are used by models in producing a particular output (Elazar et al., 2021; Belinkov, 2022; Davies et al., 2023) and characterize how these representations are leveraged in producing particular behaviors (Elhage et al., 2021; Olsson et al., 2022; Conmy et al., 2023).

Beyond simply interpreting LLMs' representation and use of information about learners, it is also important to utilize counterfactual explanation techniques to predict how their behaviors will change in response to different input prompts (Wachter et al., 2017; Ribeiro et al., 2020)—for instance, if we add a minor typo to a student's essay that is otherwise exemplary, will the model provide a substantially lower assessment of the student's knowledge in response? Conversely, it is equally important to characterize when and how models will remain invariant with respect to a given input property (Schwab and Karlen, 2019)—for example, it may be important to understand whether models always provide the same answer to different ways of phrasing the same question, meaning that they are *invariant* with respect to question semantics. Knowing the set of properties to which a given model is invariant allows us to predict whether its behavior will remain consistent if those same properties are held constant, even as other input properties may vary (Peters et al., 2016; Arjovsky et al., 2019). Between these two lines of research, we can build a systematic picture of when, how, and why model behaviors are expected to change (under counterfactuals) or remain the same (given invariances).

3.2 Whole learner support through adaptive technologies

Our vision for personalized learning that supports whole-learner adaptation necessitates a dynamic approach to learner modeling, capable of capturing and integrating the learner's complex states and needs. High-quality adaptive feedback is contingent on an accurate representation of the learner. Crafting and updating this representation is the job of the learner model, which is typically considered a separate component from the adaptation module that produces feedback in ITSs and PLEs (Shute and Zapata-Rivera, 2008). While learner models can come in many forms, such as cognitive models, machine learning models, or Bayesian networks, GenAI models like LLMs are beginning to be tested for this task (Zhang et al., 2024).

Integrating whole learner models with LLM-based support involves using cognitive, affective, or behavioral states from learner models as inputs to the adaptation module or dialogue engine (Zapata-Rivera and Forsyth, 2022). To capture the whole learner, multiple traditional models representing distinct aspects of the learner can either form a larger learner *module* or be integrated into a holistic model. Alternatively, a single LLM might serve as both the learner model and the adaptation module, though the current lack of LLM interpretability challenges trustworthiness and validation. Another viable option is leveraging a LLM to integrate outputs from various traditional learner models, providing a comprehensive inference to the adaptation module. This approach could be very useful, despite the limited research on integrating diverse types of learner information, as it offers a more nuanced understanding of the student.

Regardless of the specific system architecture used, LLMs enable just-in-time adaptive conversational feedback. This allows conversational complexity to adjust dynamically based on the learner's real-time progress, maintaining an appropriate level of challenge and promoting engagement (Zapata-Rivera and Forsyth, 2022). By basing this feedback on a rich understanding of the learner from the learner model, it offers whole-learner adaptation, potentially providing more nuanced, personalized support than existing PLEs.

3.3 Authoring and evaluating agents

When building agents to support the whole learner, the ability to operationalize a given theoretical model or dynamically incorporate new developments from learning sciences into agentic behavior "on the fly" is a desirable trait, helping to avoid the tedious process of manually re-authoring the agents. Efficient attention mechanisms (Shen et al., 2021), attention-alternatives (Gu and Dao, 2023), techniques such as retrieval-augmented generation (RAG) (Lewis et al., 2020), and needle-in-the-haystack capabilities (Kuratov et al., 2024) will enable authors to quickly reshape agent behavior, potentially even allowing them to directly operationalize longer documents such as scientific reviews or books describing evidence-based practices.

Equally important to authoring is evaluating the model outputs for faithfulness and robustness. Although preliminary experimental results with using LLMs in economics and psychology suggest that LLMs are capable of accurately mimicking aspects of human behavior like decision-making (Jia et al., 2024) and personality traits (Frisch and Giulianelli, 2024), further research is needed to generalize these findings to educational settings.

Finally, authoring is not just about *designing* agents for pedagogical support, but also developing realistic testbeds to *evaluate* them. For this line of work, authoring multi-agent social simulations (see, e.g., Park et al., 2022, 2023) will be an integral component of the end-to-end development process of ITSs and PAs. Such evaluations can ensure that the agents perform well across a wide range of scenarios, increasing educator confidence. For instance, instead of testing a PA against a single learner simulation, authoring an entire classroom of LLMs comprising multiple learner agents allows for more holistic and rigorous testing, ensuring the PA is pedagogically effective, equitable, safe, and robust before being deployed to real learners.

4 Ethical considerations

Several important ethical considerations must be addressed before deploying GenAI for educational applications. First, interpretability is crucial for trustworthiness: one cannot fully trust a model in sensitive applications like education without understanding how it represents and interacts with users (Huang et al., 2020). Second, it is important to ensure that LLMs do not exacerbate the *digital divide* in education (i.e., inequitable access to educational technologies and associated benefits), as anticipated by Capraro et al. (2024). For instance, given the substantial compute required to deploy the largest and most capable LLMs, it may be helpful to develop more compute-efficient language models for use in educational settings with limited resources (Hoffmann et al., 2022); and interdisciplinary collaborations between AI research and learning sciences will be essential in ensuring that new technologies are actually improving learning outcomes and student welfare (cf. Dahlin, 2021).

Finally, perhaps most important are concerns regarding student privacy—for instance, in the adaptive support modules envisioned above, LLMs might be provided with information about learners' emotional states to provide more holistic, empathetic feedback; but in order to protect students' privacy and ensure that sensitive information about them cannot be used for non-educational purposes such as advertising, student data should only be visible to systems with robust security and data privacy guarantees (and not, e.g., included in prompts used as input to third-party AI systems, which may use such information to train future public-facing models). These concerns are particularly significant for minors, who have special legal privacy protections and may be more vulnerable to unintended GenAI behaviors.

5 Conclusion

In this paper, we explored the potential integration of LLMs into PLEs to support the whole learner addressing both cognitive and non-cognitive characteristics. Our discussion has highlighted significant opportunities as well as challenges in integrating LLMs into PLEs, focusing on developing interpretable learner representations, adaptive technologies for personalized support, and authoring and evaluating PAs. For future research, it will be important to develop methods to enhance LLM interpretability and explainability within educational settings, facilitating trustworthiness and appropriate use of student information. Additionally, LLMs' adaptability must also be refined to ensure that models can offer individualized support that accounts for diverse learner needs and backgrounds. Finally, authoring PAs will require more principled prompting protocols, including an understanding of both relevant subject matter and pedagogical best practices, in order to engender more faithful and robust agents. By advancing each of these areas, LLMs can be better positioned to fulfill their potential as transformative tools in education, making widely-accessible personalized learning a practical reality. Through all these advancements, it is essential to be mindful of the security, privacy, and ethical concerns surrounding the handling of learner data.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

AM: Conceptualization, Writing - original draft, Writing - review & editing. AD: Conceptualization, Writing - original draft, Writing - review & editing. JP: Conceptualization, Writing - original draft. SZ: Conceptualization, Writing - original draft. DO: Conceptualization, Writing - original draft. NS: Conceptualization, Writing - original draft, Supervision, Writing - review & editing. BL: Conceptualization, Writing - original draft, Writing - review & editing. DZ-R: Conceptualization, Supervision, Writing - review & editing. CZ: Conceptualization, Supervision, Writing - review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This material was based upon work supported by the National Science Foundation and the Institute of Education Sciences under Grant #2229612.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor AS-C is currently organizing a Research Topic with the author DZ-R.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission.

This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education.

References

Abyaa, A., Khalidi Idrissi, M., and Bennani, S. (2019). Learner modelling: systematic review of the literature from the last 5 years. *Educ. Technol. Res. Dev.* 67, 1105–1143. doi: 10.1007/s11423-018-09644-1

Acosta, P., Cunningham, W., and Muller, N. (2015). *Beyond Qualifications: Labor Market Returns to Cognitive Skills and Personality Traits in Urban Colombia*. Bonn: Institute for the Study of Labor (IZA).

Anghel, B., and Balart, P. (2017). Non-cognitive skills and individual earnings: new evidence from piaac. *Series* 8, 417–473. doi: 10.1007/s13209-017-0165-x

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*. doi: 10.48550/arXiv.1907.02893

Arroyo, I., Muldner, K., Burleson, W., and Woolf, B. P. (2014). Adaptive interventions to address students—negative activating and deactivating emotions during learning activities. *Des. Recommend. Intell. Tutor. Syst.* 2, 79–92.

Baker, R. S. d., Corbett, A. T., and Aleven, V. (2008). "More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing," in *Intelligent Tutoring Systems: 9th International Conference, ITS 2008, Montreal, Canada, June 23–27, 2008 Proceedings 9* (Berlin: Springer), 406–415.

Barreto, F., Moharkar, L., Shirodkar, M., Sarode, V., Gonsalves, S., and Johns, A. (2023). "Generative artificial intelligence: Opportunities and challenges of large language models," in *International Conference on Intelligent Computing and Networking* (Berlin: Springer), 545–553.

Belinkov, Y. (2022). Probing classifiers: promises, shortcomings, and advances. *Comput. Linguist.* 48, 207–219. doi: 10.1162/coli a 00422

Bernacki, M. L., Greene, M. J., and Lobczowski, N. G. (2021). A systematic review of research on personalized learning: personalized by whom, to what, how, and for what purpose(s)? *Educ. Psychol. Rev.* 33, 1675–1715. doi: 10.1007/s10648-021-09615-8

Bull, S. (2020). There are open learner models about! *IEEE Trans. Learn. Technol.* 13, 425–448. doi: 10.1109/TLT.2020.2978473

Cade, W. L., Copeland, J. L., Person, N. K., and D'Mello, S. K. (2008). "Dialogue modes in expert tutoring," in *Intelligent Tutoring Systems: 9th International Conference, ITS 2008, Montreal, Canada, June 23-27, 2008 Proceedings 9* (Berlin: Springer), 470–479.

Calvo, R. A., and D'Mello, S. K. (2011). *New Perspectives on Affect and Learning Technologies, Vol. 3*. Cham: Springer Science & Business Media.

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., et al. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *Proc. Natl. Acad. Sci. U. S. A. Nexus* 3:191. doi: 10.1093/pnasnexus/pgae191

Cinque, M., Carretero, S., and Napierala, J. (2021). *Non-cognitive Skills and Other Related Concepts: Towards a Better Understanding of Similarities and Differences. Technical report, JRC Working Papers Series on Labour, Education and Technology*. European Commission, Joint Research Centre (JRC), Seville Spain.

Conati, C., Porayska-Pomsta, K., and Mavrikis, M. (2018). AI in education needs interpretable machine learning: lessons from open learner modelling. *arXiv preprint arXiv:1807.00154*. doi: 10.48550/arXiv.1807.00154

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*. doi: 10.48550/arXiv.2304.14997

Dahlin, E. (2021). Mind the gap! on the future of AI research. *Human. Soc. Sci. Commun.* 8, 1–4. doi: 10.1057/s41599-021-00750-9

Davies, A., Jiang, J., and Zhai, C. (2023). Competence-based analysis of language models. *arXiv preprint arXiv:2303.00333*. doi: 10.48550/arXiv.2303.00333

D'Mello, S., Lehman, B., and Person, N. (2010). "Expert tutors feedback is immediate, direct, and discriminating," in *Twenty-Third International FLAIRS Conference* (Daytona Beach, FL).

Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. (2021). Amnesic probing: behavioral explanation with amnesic counterfactuals. *Trans. Assoc. Comput. Linguist.* 9, 160–175. doi: 10.1162/tacl_a_00359

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., et al. (2021). A mathematical framework for transformer circuits. *Transform. Circ. Thread* 1.

Frisch, I., and Giulianelli, M. (2024). LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *arXiv preprint arXiv:2402.02896*. doi: 10.48550/arXiv.2402.02896

Fu, Y., Zhang, M., Nguyen, L. K., Lin, Y., Michelson, R., Tayebi, T. J., et al. (2023). "Self-talk with superhero zip: supporting children's socioemotional learning with conversational agents," in *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference* (Chicago, IL), 173–186.

Gu, A., and Dao, T. (2023). Mamba: linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*. doi: 10.48550/arXiv.2312.00752

Gui, G., and Toubia, O. (2023). The challenge of using LLMs to simulate human behavior: a causal inference perspective. *arXiv preprint arXiv:2312.15524*. doi: 10.48550/arXiv.2312.15524

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., et al. (2022). "An empirical analysis of compute-optimal large language model training," in *Advances in Neural Information Processing Systems, Vol. 35*, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (Red Hook, NY: Curran Associates, Inc.), 30016–30030.

Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., et al. (2020). A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Comput. Sci. Rev.* 37:100270. doi: 10.48550/arXiv.1812.08342

Ismail, H., Hussein, N., Harous, S., and Khalil, A. (2023). Survey of personalized learning software systems: a taxonomy of environments, learning content, and user models. *Educ. Sci.* 13:741. doi: 10.3390/educsci13070741

Jia, J., Yuan, Z., Pan, J., McNamara, P., and Chen, D. (2024). *Decision-Making Behavior Evaluation Framework for LLMs Under Uncertain Context. arXiv [Preprint]. arXiv*:2406.05972.

Kambhampati, S., Valmeekam, K., Guan, L., Stechly, K., Verma, M., Bhambri, S., et al. (2024). LLMs can't plan, but can help planning in LLM-modulo frameworks. *arXiv preprint arXiv:2402.01817*. doi: 10.48550/arXiv.2402.01817

Käser, T., and Alexandron, G. (2023). Simulated learners in educational technology: a systematic literature review and a turing-like test. *Int. J. Artif. Intell. Educ.* 23:2. doi: 10.1007/s40593-023-00337-2

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., and Borghans, L. (2014). *Fostering and Measuring Skills: Improving Cognitive and Non-cognitive Skills to Promote Lifetime Success*. Cambridge, MA: National Bureau of Economic Research.

Kay, J., Zapata-Rivera, D., and Conati, C. (2020). The gift of scrutable learner models: why and how. *Des. Recommend. Intell. Tutor. Syst.* 8, 25–40.

Koedinger, K. R., Matsuda, N., MacLellan, C. J., and McLaughlin, E. A. (2015). "Methods for evaluating simulated learners: examples from Simstudent," in *17th International Conference on Artificial Intelligence in Education*, Madrid Spain.

Kuratov, Y., Bulatov, A., Anokhin, P., Sorokin, D., Sorokin, A., and Burtsev, M. (2024). In search of needles in a 10 m haystack: recurrent memory

finds what LLMs miss. *arXiv preprint arXiv:2402.10790*. doi: 10.48550/arXiv.2402.10790

Lehman, B., Sparks, J., Zapata-Rivera, D., Steinberg, J., and Forstyth, C. (2024). A framework of caring assessments for diverse learners. *Pract. Assess. Res. Eval.* 29:9. doi: 10.7275/pare.2102

Lepper, M. R., Woolverton, M., Mumme, D. L., and Gurtner, J.-L. (1993). *Motivational Techniques of Expert Human Tutors: Lessons for the Design of Computer-Based Tutors*. Routledge.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inform. Process. Syst.* 33, 9459–9474. doi: 10.48550/arXiv.2005.11401

Liao, Q. V., and Vaughan, J. W. (2023). AI transparency in the age of LLMs: a human-centered research roadmap. *arXiv preprint arXiv:2306.01941*. doi: 10.48550/arXiv.2306.01941

Liu, V., and Chilton, L. B. (2022). "Design guidelines for prompt engineering text-to-image generative models," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22* (New York, NY: Association for Computing Machinery), 1–23.

Liu, Z., Duan, H., Liu, S., Mu, R., Liu, S., and Yang, Z. (2024). Improving knowledge gain and emotional experience in online learning with knowledge and emotional scaffolding-based conversational agent. *Educ. Technol. Soc.* 27, 197–219. doi: 10.30191/ETS.202404_27(2).RP08

Loya, M., Sinha, D., and Futrell, R. (2023). "Exploring the sensitivity of LLMs' decision-making capabilities: insights from prompt variations and hyperparameters," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, eds. H. Bouamor, J. Pino, and K. Bali (Singapore: Association for Computational Linguistics), 3711–3716.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. (2022). "Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. S. Muresan, P. Nakov, and A. Villavicencio (Dublin: Association for Computational Linguistics), 8086–8098.

Lu, Y., Wang, D., Meng, Q., and Chen, P. (2020). "Towards interpretable deep learning models for knowledge tracing," in *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21* (Berlin: Springer), 185–190.

Mannekote, A., Celepkolu, M., Wiggins, J. B., and Boyer, K. E. (2023). "Exploring usability issues in instruction-based and schema-based authoring of task-oriented dialogue agents," in *Proceedings of the 5th International Conference on Conversational User Interfaces, CUI '23* (New York, NY: Association for Computing Machinery), 1–6.

McQuiggan, S. W., Robison, J. L., Phillips, R., and Lester, J. C. (2008). "Modeling parallel and reactive empathy in virtual agents: an inductive approach," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems-Vol. 1* (Estoril: Citeseer), 167–174.

Mercado, F. (2018). Whole child framework: supporting educators in their plight toward mtss and equity. *J. Leaders. Eq. Res.* 4.

Miliv cka, J., Marklová, A., Van Slambrouck, K., Pospiv silová, E., Simsová, J., Harvan, S., et al. (2024). Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *PLoS ONE* 19:e0298522. doi: 10.1371/journal.pone.0298522

Mohammadi, B. (2024). Wait, it's all token noise? always has been: interpreting LLM behavior using Shapley value. *arXiv:2404.01332 [cs]*. doi: 10.48550/arXiv.2404.01332

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., et al. (2022). In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*. doi: 10.48550/arXiv.2209.11895

Oppenlaender, J., Linder, R., and Silvennoinen, J. (2023). Prompting AI art: an investigation into the creative skill of prompt engineering. *arXiv:2303.13534 [cs]*. doi: 10.48550/arXiv.2303.13534

Ortega-Ochoa, E., Arguedas, M., and Daradoumis, T. (2024). Empathic pedagogical conversational agents: a systematic literature review. *Br. J. Educ. Technol.* 55, 886–909. doi: 10.1111/bjet.13413

Pardos, Z. A., and Bhandari, S. (2023). Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint arXiv:2302.06871*. doi: 10.48550/arXiv.2302.06871

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: interactive simulacra of human behavior. *arXiv:2304.03442 [cs]*. doi: 10.48550/arXiv.2304.03442

Park, J. S., Popowski, L., Cai, C., Morris, M. R., Liang, P., and Bernstein, M. S. (2022). "Social simulacra: creating populated prototypes for social computing systems," in *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR), 1–18.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *J. Royal Stat. Soc. Ser. B* 78, 947–1012. doi: 10.48550/arXiv.1501.01332

Pinto, J. D., Paquette, L., and Bosch, N. (2023). "Interpretable neural networks vs. expert-defined models for learner behavior detection," in *Companion Proceedings of the 13th International Conference on Learning Analytics and Knowledge Conference (LAK23)* (Arlington, TX), 105–107.

Plass, J. L., and Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *J. Res. Technol. Educ.* 52, 275–300. doi: 10.1080/15391523.2020.1719943

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: behavioral testing of NLP models with checklist. *arXiv preprint arXiv:2005.04118*. doi: 10.48550/arXiv.2005.04118

Sabourin, J., Mott, B., and Lester, J. (2011). Computational models of affect and empathy for pedagogical virtual agents. *Stand. Emot. Model.* 2011, 1–14.

Schwab, P., and Karlen, W. (2019). "CXPlain: causal explanations for model interpretation under uncertainty," in *Advances in Neural Information Processing Systems, Vol. 32*, eds. H. Wallach, H. Larochelle, A. Beygelzimer, F. dsingle Alché-Buc, E. Fox, and R. Garnett (New York, NY: Curran Associates, Inc.), 19.

Shen, Z., Zhang, M., Zhao, H., Yi, S., and Li, H. (2021). "Efficient attention: attention with linear complexities," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Waikoloa, HI), 3531–3539.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. *Int. J. Hum. Comput. Stud.* 146:102551. doi: 10.1016/j.ijhcs.2020.102551

Shute, V. J., and Zapata-Rivera, D. (2008). "Adaptive technologies," in *Handbook of Research on Educational Communications and Technology* (London: Routledge), 277–294.

Sottilare, R., Graesser, A., Hu, X., and Brawner, K. (2015). *Design Recommendations for Intelligent Tutoring Systems - Volume 3: Authoring Tools and Expert Modeling Techniques*. Memphis, TN.

Swamy, V., Blackwell, J., Frej, J., Jaggi, M., and Käser, T. (2024). InterpretCC: conditional computation for inherently interpretable neural networks. *arXiv preprint arXiv:2402.02933*. doi: 10.48550/arXiv.2402.02933

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., et al. (2024). Scaling monosemanticity: extracting interpretable features from claude 3 sonnet. *Transform. Circ. Thread*. Available online at: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL Tech.* 31:841. doi: 10.48550/arXiv.1711.00399

Walkington, C., and Bernacki, M. L. (2018). Personalization of instruction: design dimensions and implications for cognition. *J. Exp. Educ.* 86, 50–68. doi: 10.1080/00220973.2017.1380590

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., et al. (2023). "Self-instruct: aligning language models with self-generated instructions," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. A. Rogers, J. Boyd-Graber, and N. Okazaki (Toronto, ON: Association for Computational Linguistics), 13484–13508.

Weitekamp, D., Rachatasumrit, N., Wei, R., Harpstead, E., and Koedinger, K. (2023). "Simulating learning from language and examples," in *International Conference on Artificial Intelligence in Education* (Berlin: Springer), 580–586.

Yen, W. M., and Fitzpatrick, A. R. (2006). Item response theory. *Educ. Measur.* 4, 111–153. Available online at: https://intelligence.org/files/AlignmentHardStart.pdf

Yudkowsky, E. (2016). The AI alignment problem: why it is hard, and where to start. *Symbol. Syst. Distinguis. Speak.* 4:1.

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., and Yang, Q. (2023). "Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg), 1–21.

Zapata-Rivera, D., and Arslan, B. (2021). "Enhancing personalization by integrating top-down and bottom-up approaches to learner modeling," in *International Conference on Human-Computer Interaction* (Berlin: Springer), 234–246.

Zapata-Rivera, D., and Forsyth, C. M. (2022). "Learner modeling in conversation-based assessment," in *International Conference on Human-Computer Interaction* (Berlin: Springer), 73–83.

Zapata-Rivera, J.-D., and Greer, J. E. (2004). Interacting with inspectable bayesian student models. *Int. J. Artif. Intell. Educ.* 14, 127–163. doi: 10.5555/1434858.1434859

Zhang, L., Lin, J., Borchers, C., Sabatini, J., Hollander, J., Cao, M., et al. (2024). "Predicting learning performance with large language models: a study in adult literacy," in *Adaptive Instructional Systems*, eds. R. A. Sottilare and J. Schwarz (Cham: Springer Nature Switzerland), 333–353.