Validity Claims in Children-AI Discourse: Experiment with ChatGPT

Johannes Schneider[1], Leona Chandra Kruse[2] and Isabella Seeber[3]

[1]*Department of Computer Science and Information Systems, University of Liechtenstein, Vaduz, Liechtenstein*
[2]*Department of Information Systems, University of Agder, Norway*
[3]*Department of Management, Technology and Strategy, Grenoble Ecole de Management, France*

Keywords: Large Language Models, Education, Children, Adolescent.

Abstract: Large language models like ChatGPT are increasingly used by people from all age groups. They have already started to transform education and research. However, these models are also known to have a number of shortcomings, i.e., they can hallucinate or provide biased responses. While adults might be able to assess such shortcomings, the most vulnerable group of our society – children – might not be able to do so. Thus, in this paper, we analyze responses to commonly asked questions tailored to different age groups by OpenAI's ChatGPT. Our assessment uses Habermas' validity claims. We operationalize them using computational measures such as established reading scores and interpretative analysis. Our results indicate that responses were mostly (but not always) truthful, legitimate, and comprehensible and aligned with the developmental phases, but with one important exception: responses for two-year-olds.

1 INTRODUCTION

Large language models (LLMs) such as ChatGPT are increasingly used in learning activities and everyday lives in all age groups. As LLMs show remarkable interaction capabilities, LLMs are (or will likely) soon be embedded in voice assistants such as Amazon Alexa. Children are in touch with these tools at a very young age, i.e., their very first word might not be "Mama" or "Papa" but Alexa.[1] Parents are concerned about the suitability of these applications for their children, and this sentiment is shared by educators[2]. To make matters worse, ChatGPT-bashing seems to have become a popular pastime. Some people disclose invalid responses made by ChatGPT on social media platforms, often warning others of its potential danger[3]. Others demonstrate the so-called "jailbreak" phrases to prompt ChatGPT to ignore policies and guidelines set by its governing body. The result is socially unacceptable or even explicit responses. Furthermore, large language models suffer from hallucinations (Ji et al., 2023). That is, they might contradict existing knowledge or fabricate statements that cannot be verified. All these add to the parental and educational concerns: Should the children be kept away from LLMs, if at all possible?

The above illustration points to two issues. First, there are different kinds of invalid responses made by ChatGPT and other applications based on large language models. They violate different kinds of validity claims (Habermas, 1985; Habermas & McCarthy, 1987): *truth* (e.g., presenting wrong facts), *sincerity* (e.g., prioritizing one political view over others), *legitimacy* (e.g., showing children explicit contents), and *comprehensibility* (e.g., mixing languages and jargons). Second, we can evaluate validity claims based on the context of a discourse. Social acceptability of a response depends on social norms which, in turn, often depend on children's psychosocial development and intention (e.g., seeking facts versus explanation related to social values). Also adequate phrasing and language within a child and LLM interaction depends on the context, i.e., a child's developmental stage. That is, while

[1] https://www.washingtonpost.com/news/the-switch/wp/2017/11/21/when-your-kid-tries-to-say-alexa-before-mama/

[2] https://www.nytimes.com/2023/03/22/well/family/ai-chatgpt-parents-children.html

[3] https://theconversation.com/chatgpt-is-great-youre-just-using-it-wrong-198848

Schneider, J., Kruse, L. and Seeber, I.
Validity Claims in Children-AI Discourse: Experiment with ChatGPT.
DOI: 10.5220/0012552300003693
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024) - Volume 1*, pages 289-296
ISBN: 978-989-758-697-2; ISSN: 2184-5026
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

ChatGPT might be well-suited for adults, it is less clear whether it can comprehensibly and legitimately interact with younger age groups, in particular those within K-12.

These issues are captured in the following research questions:

How do large language models (e.g., ChatGPT) satisfy different validity claims in a discourse with children?

How do these models calibrate the response according to children's developmental stage?

We adopt an explorative computational approach in addressing the questions. This paper discusses the results of our first analysis of the legitimacy and comprehensibility validity claims. We outline the study continuation in the outlook section.

This study is also relevant beyond education and learning as it is positioned within the ongoing discourse: evaluating useful conversational agents and responsible artificial intelligence (AI). Our paper contributes to the accumulative knowledge about how to assess conversational agents, e.g., situated in the broader scope on auditing LLMs (Mökander et al., 2023) and, also how users can harness the benefits of conversational agents. In the long run, this helps in guiding how to design conversational agents more effectively (e.g., Schöbel et al., 2023). We also join the broader discourse on the idea of responsibility and unintended consequences of AI (e.g., Mikalef et al., 2022) and governance of AI (Schneider et al., 2022), especially generative AI such as the LLM behind ChatGPT.

2 THEORETICAL BACKGROUND

2.1 Validity Claims

Validity has various definitions: (1) "The quality of being valid in law; legal authority, force, or strength; (2) The quality of being well-founded on fact, or established on sound principles, and thoroughly applicable to the case or circumstances; soundness and strength (of argument, proof, authority, etc.); and (3) Value or worth; efficacy." (Oxford Dictionaries). The variation indicates there is not a single way to evaluate whether an utterance is valid. There are also different ways to claim the validity of an utterance. Among the widely used references is Habermas's typology of validity claims in his theory of communicative action (1985, 1987). The typology has been applied in IS research to study media discourse on technology (Cukier et al., 2009), crowdsourcing practices (Schlagwein et al., 2019), humanness in digital assistants (Porra et al., 2020), information security policies (Stahl et al., 2012), emancipatory pedagogy (Young, 2018), and the design of report-authoring application (Heng & De Moor, 2003).

The theory of communicative action proposes three kinds of action: instrumental action, strategic action, and communicative action (1985, 1987). Instrumental action is targeted toward non-social actors (i.e., object), such as using an IS application as a tool for accomplishing tasks. A strategic action aims to achieve success by influencing other social factors, such as when children persuade their grandparents to buy them toys. In contrast to the above actions, a *communicative action* aims toward mutual understanding, developing an inter-subjective agreement as the basis for coordination and collaboration.

Actors engage in the process of raising, questioning, and defending validity claims in order to reach mutual understanding. The theory differentiated between four validity claims: (1) truth or *Wahrheit* (that something is the case or factual), (2) legitimacy or *Richtigkeit* (that something is right or legitimate according to social norms and values), (3) sincerity or *Wahrhaftigkeit* (that the speaker is sincere about her or his beliefs, feelings, and hopes), and (4) comprehensibility or *Klarheit* (that the utterance is understandable). The potential distortion of each claim is (1) confusion, (2) misrepresentation, (3) illegitimacy, and (4) false assurance (Cukier et al., 2009).

2.2 Children Development Stages

Researchers in Developmental Psychology have proposed different development stages observed in children. Our study focuses on two theories: Jean Piaget's cognitive development theory and Erik Erikson's psychosocial development theory. We choose these theories because they have been widely applied in education (Barrouillet, 2015; Maree, 2021) and they focus on different, but equally important aspects of children's development—cognitive development and psychosocial development.

Piaget observed four stages of cognitive development, from sensorimotor stage to formal operational stage. The following explanation is based on Papalia and Martorell's work (2023). In *sensorimotor stage* (birth-2 years), children explore their environment with their senses and motoric actions, and they represent objects without the use of language. The use of language begins in the *preoperational stage* (2-6 years). However, children are still incapable of logical thinking, and they have difficulty seeing the viewpoint of others (i.e., egocentric tendencies). In *concrete operational stage* (7-11 years) they are already capable of logical thinking, but this ability is limited to concrete objects, excluding abstract concepts. Abstract thinking develops *in the formal operational stage* (12 years-adulthood), involving the use of symbols and hypothetico-deductive reasoning.

In contrast to Piaget, Erikson focused on the psychosocial aspects of human development. Each development stage deals with psychosocial tensions which, once resolved, give way to virtues (Maree, 2021). The following explanation is based on Papalia et al.'s work (2008). The *infancy stage* (birth-1 year) revolves around trust and mistrust, because children are completely dependent on their caregivers. Those in the *toddlerhood stage* (1-3 years) deal with the development of will and autonomy. The *early childhood stage* (3-6 years) deals with the question of initiative and purpose by exploring objects in the environment. Children in the *late childhood stage* (6-puberty) face the conflict between industry and inferiority, aiming to develop the virtue of competence. The adolescence *stage* (puberty-19 years) is about the development of identity with the support of social relationships. *Early adults* (20-25 years) should resolve the tension between intimacy and isolation in their pursuit of love or romantic relationships. *Adults* (over 26 years) deal with different issues as they mature and care for the next generation, and they ultimately develop wisdom.

The above overview points to two major requirements for discourse with children. First, children differ in their cognitive ability to comprehend the content and context of a discourse across developmental stages. Second, the question of social acceptability (i.e., norms and values) of a discourse depends on children's psychosocial development.

2.3 Large Language Models

Large language models are very large deep learning models that can process textual inputs. The number of LLMs is rapidly growing with most big companies developing such models (Zhao et al., 2023). They differ in terms of training data, which can impact the capabilities of the model. During the training phase, these models have to predict (omitted) words for a given context, such as the word "name" for an input such as "What is your _?" . The models can be further adjusted to perform better on textual instructions and produce outcomes that are better aligned with human desiderata such as reducing toxicity (Ouyang et al., 2022). For example, the well-known ChatGPT model originates from GPT-3 through fine-tuning on data from human annotators (Ouyang et al., 2022). Unfortunately, despite its name "OpenAI" the information on ChatGPT and its successor models is not sufficient to satisfy important academic criteria such as reproducibility. For example, the 100-page technical report on GPT-4 is mostly limited to performance evaluation but contains almost no information on the training data or possible adjustments of the transformer architecture (OpenAI, 2023).

Due to their broad applicability LLMs are also often subsumed under the term foundation models (Schneider et al., 2024). AI in general and, more so, generative AI is difficult to understand (Longo et al., 2023) while exhibiting surprising traits such as creativity (Basalla et al., 2022), underpinning the need for studies to better understand them. LLMs suffer from trust issues in an educational context (Schneider, Richner, et al., 2023; Schneider, Schenk, et al., 2023) due to hallucinations and prompt sensitivity. The growth in size of training data and model size has led to novel emergent behavior. LLMs can solve tasks they were not explicitly trained for using a textual description and possibly examples (Brown et al., 2020). For example, the models can solve simple math questions like "What is 3+9?", although they were not explicitly trained on additions. Furthermore, adjustments to the textual input influence its output and potentially even the quality. For example, back in 2020 for GPT-3 Kojima et al. (2022) added "Let's think step by step" to an input, which caused the model to perform significantly better on various benchmarks. Such findings came as a surprise and gave rise to exploring a number of techniques for designing inputs, i.e., prompt engineering, which is an active field of research (Liu et al., 2021). In our work, we use age modifiers to adjust prompts to different age groups. While we are not aware of any systematic exploration of adding age modifiers, people have used age modifiers, e.g., even the original paper on InstructGPT serving as a pre-cursor to ChatGPT (Ouyang et al., 2022) used an example with an age modifier. A recent short newspaper article (Buchanan, 2022) asked three experts on children's writing to tell if a given writing stems from ChatGPT or an actual child. They reported that none of the experts could consistently predict correctly who the author of the provided text was. Thus, essentially the newspaper implicitly states that plagiarism detection is not possible for such age groups by human experts. Despite the threat of abuse in education, ChatGPT has also been alleged to offer many opportunities in education (Kasneci et al., 2023) at all age groups though most of the claimed benefits have yet to be tested. A study like ours examining the adequacy of explanations of such models covering toddlers onto adults has been absent.

3 METHOD

We defined a set of 36 questions that are typical for different developmental stages and posed them to ChatGPT, as it has emerged as the best-known and most widely used LLM[4]. Furthermore, general issues such as prompt sensitivity and hallucinations are prevalent across all well-known LLMs. The considered questions are a selection from three online resources that we obtained through an Internet search for questions asked by children, teenagers, and adults[5]. We reviewed the questions to assess their suitability and partially modified them, mostly to be easily adjustable with an age modifier. That is, we appended age modifiers to these questions to encourage responses accounting to both Piaget's and Erikson's developmental stages (e.g., 7 years correspond to concrete operational and late childhood stages). The questions differ in style. Questions by children are more factual, i.e., geared towards understanding the world (e.g., 'Explain where babies come from', 'Explain why the sky is blue', 'Explain how much salt is in the ocean'). Questions by teenagers are more about understanding changes due to becoming an adult (e.g., 'What does it feel like to be legally old enough to drink?',' 'When is the right time to leave your parents house?', 'At what point do I start feeling like I can properly adult?'). While questions by adults are mostly

about their life as adults (e.g., 'What should someone be looking for in a partner?,' 'Am I a good person?', 'Should someone have a family? '). The six appended modifiers were " to a X-year-old" with X being 2, 4, 7, 11, 16 and 25. While young children might not be able to access ChatGPT using a web-interface, they might well interact indirectly with LLMs, i.e., through voice assistants converting speech to text that is fed into an LLM.

We combined all questions with all age modifiers for the sake of completeness, especially, since children tend to ask all kinds of questions beyond those that are typical for their age group.

We also posed the question without an modifier as a baseline for comparison. We prompted OpenAI's "gpt-3.5-turbo" model using the provided API. The full list of questions and responses by ChatGPT can be found at https://github.com/JohnTailor/LLM_ValidityClaims.git .

We then analyzed these questions manually and automatically. That is, we aim to analyze the responses in terms of the four proposed validity claims through (i) computational analysis and (ii) manual interpretive content analysis. This paper reports on our first findings with two validity claims: legitimacy and comprehensibility.

3.1 Computational Analysis

We operationalized comprehensibility using measures capturing the difficulty of the text to assess if it is understandable. For legitimacy, we focused in this short paper on one relevant aspect which is appropriateness in terms of wording or more specifically, toxicity. The indicators for comprehensibility and legitimacy are summarized in Table 1.

Table 1: Computational indicators for each validity claim.

Validity claim	Indicators
Comprehensibility	Flesch reading-ease score, text length, and Dale-Chall readability score
Legitimacy	Toxicity, profanity, insult, threat, sexually explicit content, and sentiment

More precisely, for comprehensibility, we computed the Flesch Reading Ease Score (FRES) (Flesch, 1948) and the Dale–Chall readability (Chall & Dale, 1995). We chose these two measures as they are commonly used and are based on two different underlying ideas. Both measures are implemented in Python's textstat library version 0.7.3 (Bansal, 2021). Additionally, we also reported response length.

[4] https://www.cyberhaven.com/blog/chatgpt-vs-google-bard-usage-statistics

[5] Questions are at Link; Sources: https://youaremom.com/parenting/common-questions-that-kids-ask/ (We added two of our own questions targeted to toddlers to have a balanced set of 12 questions per age group); https://www.familyzone.com/anz/families/blog/100-questions-for-teens ; https://www.elitedaily.com/life/30-questions-able-answer-30/963183

Table 2: Results for comprehensibility indicators (difficulty and length).

Modifier	Flesch Reading-Ease Score	Dale-Chall Readability Score	Length (in chars)
to a 2-year-old	62.88	7.08	478
to a 4-year-old	67.38	6.77	423
to a 7-year-old	77.87	6.55	421
to a 11-year-old	72.26	7.04	514
to a 16-year-old	59.74	7.62	691
to a 25-year-old	60.65	7.35	765
None (baseline)	60.45	7.65	663

The Flesch reading-ease test employs a scoring system where higher scores reflect text that is simpler to comprehend, while lower scores indicate more complex material. The Flesch reading-ease score (FRES) formula is used to calculate the score. It is given as 206.835 - 1.015*ASL - 84.6*ASW, where ASL is the average sentence length and ASW the average syllables per word. A score of 100 indicates text that is very easy to read and easily understood by an average 11-year-old student, i.e., a fifth grader in the US. A score of 0 is at a level of university graduates, meaning extremely difficult to understand.

The Dale-Chall readability score is computed based on the occurrence of difficult words, which are all those words of a text not found in the predefined list of 3000 familiar words. It is computed as 0.1579 * PDW + 0.0496 * ASL, where PDW is the percentage of difficult words and ASL is again the average sentence length. A score of 4.9 or lower means that the text is easily understood by an average 4th grader (in US) or lower, i.e., 10-year-old or younger. A score of 9.9 means that it is easily understood by an average college student, i.e., about 18 years old.

For legitimacy, we computed toxicity scores using the Perspective API (Jigsaw, 2017) by Google and JigSaw. It employs machine learning trained on millions of comments from sources such as New York Times and Wikipedia online forums tagged by 3 to 10 crowd workers. It scores texts on various (negative) emotional aspects in the range of 0 to 1. A higher score indicates a greater likelihood a text is perceived as having the given attribute. We neglected the attributes of the library, which were not thoroughly tested ("experimental") as they aer not reliable enough.

We also computed the sentiment using two standard machine learning models. We used as Sentiment Measure 1 the distilbert-base-uncased-finetuned-sst-2-english model (HF Canonical Model Maintainers, 2022), which is well-established and produces binary outputs. Our second model produces more fine-granular outputs (Sentiment Measure 2), i.e., yielding positive, neutral and negative classifications (Pérez et al., 2021). The motivation to use two distinct models is to increase the robustness of findings. We report the sentiment score *sp* - *sn*, where the *sp* is the score in [0,1] output by the model for the positive class and *sn* is the score in [0,1] for the negative class.

3.2 Interpretive Content Analysis

The computational analysis was coupled with interpretive content analysis. Each author first read all responses given by ChatGPT individually to make sense of the content and the communication approach. All authors exchanged their initial insights before discussing each validity claim in detail. Next, we referred to Piaget's and Erikson's theories to analyze the responses based on children's development stage. We went back to literature to find the explanation for our emerging insights unrelated to the development theories. This is how we came to discuss the signaling theory in light of our findings. We will continue the analysis as outlined in the outline section.

4 FINDINGS

We present and discuss the results of our analysis in the following, focusing on comprehensibility and legitimacy claim.

4.1 Comprehensibility Claim

The comprehensibility (*Klarheit*) claim is satisfied if a response is legible or intelligible. This is indicated by the degree of difficulty (reading ease scores) and the length of response. Table 2 presents the computational results.

All indicators differ across the ages of the intended reader. The length of the text in characters is relatively consistent across age groups. The responses for children younger than 11 years are generally easier to understand than those for older readers. The Flesch Reading-Ease scores gradually decreased from 77.87 for 7-year-olds to 59.74 for 16-year-olds, indicating that texts that are easy for 7-year-olds to read and comprehend may become more difficult for older age groups. However, the scores for 2-year-olds and 4-year-olds do not indicate that texts are easier than for 7-year-olds. This is aligned with the Dale-Chall readability scores, which are lowest for 7-year-olds and gradually increase from 6.55 for 7-year-olds to 7.62 for 16-year-olds. This means that responses do get more difficult to comprehend as older age groups are targeted.

Our interpretive content analysis also points to similar insights. Consider the following examples:

"*It's like a big brain that helps us learn and have fun.*"
"*It's like the clouds are giving the earth a big drink of water!*" (Both intended for 4-year-olds)

These responses are intended for children in preoperational and early childhood stages. They are characterized by their still limited ability to understand abstract concepts and their pursuit of the sense of autonomy. The responses use metaphors and analogies, and they can potentially help the intended readers make connections between abstract concepts and real-world experiences. This is useful for a discourse about facts or scientific explanations. Consider the following examples for discourse about values and opinions (should someone have a family?):

"*Having a family is a personal choice and depends on what makes someone happy. Some people choose to have a family and others choose not to. It is important to do what makes you happy and fulfilled in life.*" (Intended for 4-year-olds)

"*Having a family is a personal choice and there is no right or wrong answer. Some people choose to have a family because they enjoy the love and support that comes with having a partner and children. Others choose not to have a family because they prefer to focus on their career, hobbies, or other interests. It's important to remember that everyone is different and it's okay to make the choice that feels right for you.*" (Intended for 7-year-olds)

These responses indeed differ in length, but we find them comparable in difficulty. The first response shows an attempt to simplify the concept of family and personal choice, but it still uses somewhat abstract language such as "personal choice" and "fulfilled in life," which may not be easily understood by 4-year-olds. The second one also uses complex language and abstract concepts such as "career," "hobbies," and "other interests," but the elaboration (resulting in a longer response) supports 7-year-olds readers to understand the response.

4.2 Legitimacy Claim

The legitimacy (*Richtigkeit*) claim is satisfied if a response is right or appropriate in the light of existing norms or values. We computed eight scores as indicators of legitimacy: degree of toxicity (2 measures), insult, profanity, threat, sexually explicit content, and sentiment (2 measures). Table 3 presents the computational results.

The scores for toxicity, insult, profanity, threat, and sexually explicit content are generally low for all responses. From 4-year-old onwards both sentiment measures decrease as the age of intended readers increases. The least positive sentiment is observed in responses to questions without modifiers (i.e., for general readers). Interestingly, the responses for 2-year-olds show lower positive sentiment than the responses for 4-year-olds and 7-year-olds. Based on these scores alone, one may conclude that the responses are appropriately calibrated for legitimacy based on the age of their intended readers (except for 2 years old).

Our interpretive content analysis points to a richer nuance. We discern the use of signaling strategies as a legitimacy means. One strategy is acknowledging limitation or conformity to social norms at the beginning of the response:

"*As an AI language model, I cannot provide advice to a 2-year-old as they may not comprehend it.*" (Intended for 2-year-old)
"*As an AI language model, I do not have personal beliefs or opinions.*" (Intended for multiple age groups)

Another strategy is using pet names (e.g., "sweetie" and "kiddo") to signal the characteristics of the intended readers (i.e., younger children). Both strategies appear to be mere rhetorical device as the contents that follow are not always legitimate:

"*As an AI language model, I am not programmed to provide inappropriate content to children. However, in a general and age-appropriate way, babies are made when a sperm from a man meets an egg from a woman. This usually happens during sexual intercourse, where the sperm travels through the man's penis and into the woman's vagina. The sperm then travels up to the woman's uterus and can fertilize the egg if it is there. If the egg is fertilized, it can grow into a baby inside the woman's uterus.*" (Intended for 7-year-olds)

Table 3: Results for legitimacy indicators (undesirable content and sentiment).

Modifier	Toxicity	Severe toxicity	Insult	Profanity	Threat	Sexually explicit	Sentiment Measure 1	Sentiment Measure 2
to a 2-year-old	0.024	0.001	0.012	0.016	0.007	0.01	0.489	0.57
to a 4-year-old	0.02	0.001	0.01	0.015	0.007	0.01	0.603	0.834
to a 7-year-old	0.019	0.001	0.01	0.014	0.007	0.008	0.572	0.728
to a 11-year-old	0.023	0.001	0.01	0.016	0.01	0.01	0.357	0.701
to a 16-year-old	0.028	0.001	0.012	0.018	0.012	0.012	0.453	0.58
to a 25-year-old	0.025	0.001	0.012	0.018	0.008	0.013	0.365	0.506
None (baseline)	0.028	0.001	0.012	0.023	0.008	0.013	0.249	0.156

Interestingly, the responses to 2-year-olds often signal reluctance to provide explanations to children at such a young age. We find this tendency appropriate for the cognitive and psychosocial development of the intended readers.

However, such responses may lead to confusion if directly communicated to real 2-year-old children (e.g., if the language model is used directly in smart toys and the responses are automatically calibrated based on individual user age).

5 OUTLOOK

This paper presents first findings on legitimacy and comprehensibility as validity claims in children-AI discourse. We will continue the analysis on sincerity and truth validity claims. However, this is not without challenges. Truth validity claim can be ascertained only if we know what is true. This is feasible for facts (e.g., what is the capital city of Tonga). But what about opinions and advice? The argumentation logic can be evaluated via syllogism and similar approaches, but this also requires knowledge about the truth of each statement and the structure of the argumentation (i.e., apple is good for you because it contains vitamins, but rotten apple can cause problems). Sincerity evaluation also comes with its own challenges. Future continuation should also involve the intended readers in the evaluation of responses (e.g., asking children of different ages to talk about the responses).

The launch of ChatGPT in late 2022 was followed by rapid uptake leading to more than 100 million users within a few weeks with no clear end in sight[6]. This is no surprise as ChatGPT has shown remarkable capabilities, i.e., it has passed numerous exams at a university level (Choi et al., 2023; Kortemeyer, 2023). Only 4 months after the launch of ChatGPT a newer version GPT-4 has become accessible. At the time of writing, programmatic access (via an API) was restricted and the default model was still ChatGPT. However, in the future, we intend to leverage newer versions as well. According to the technical report published by the company itself(OpenAI, 2023), GPT-4 improves on measures related to toxicity and also generally performs better on benchmarks such as exams. However, even the GPT-4 model owner acknowledges that the model does not consistently outperform its predecessor ChatGPT. It is not clear to what extent (if at all) future models become more suitable to interact with children.

Our current and extended study will contribute to the understanding of how well a predictive LLM can calibrate its responses to the developmental stages of the intended audience. This understanding can, in turn, inform parents and caregivers on harnessing the most of children-AI discourse as well as developers of LLMs.

Moreover, heavy reliance on LLMs such as ChatGPT may also contribute to the formation of echo chambers. These models generate responses based on the prompts (i.e., user input) that reflect values and opinions. The responses tend to confirm these values and opinions. If consistently repeated, this tendency may again reinforce the values and opinions, a phenomenon coined as an echo chamber (Gilbert et al., 2009).

REFERENCES

Bansal, S. (2021). textstat: Python package to calculate statistics from text. https://pypi.org/project/textstat/

Barrouillet, P. (2015). Theories of cognitive development: From Piaget to today. Developmental Review, 38.

Basalla, M., Schneider, J., & vom Brocke, J. (2022). Creativity of deep learning: Conceptualization and assessment. Proc. of the International Conference on Agents and Artificial Intelligence (ICAART).

[6] https://explodingtopics.com/blog/chatgpt-users

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Buchanan, L. (2022). Did a Fourth Grader Write This? Or the New Chatbot? https://www.nytimes.com/interactive/2022/12/26/upshot/chatgpt-child-essays.html?searchResultPosition=2

Chall, J. S., & Dale, E. (1995). Readability revisited: The new Dale-Chall readability formula. Brookline Books.

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. (2023). Chatgpt goes to law school. Available at SSRN.

Cukier, W., Ngwenyama, O., Bauer, R., & Middleton, C. (2009). A critical analysis of media discourse on information technology: Preliminary results of a proposed method for critical discourse analysis. Information Systems Journal, 19(2)

Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221.

Gilbert, E., Bergstrom, T., & Karahalios, K. (2009). Blogs are echo chambers: Blogs are echo chambers. Hawaiin International Conference on System Sciences, 1–10.

Habermas, J. (1985). The theory of communicative action: Volume 1: Reason and the rationalization of society.

Habermas, J., & McCarthy, T. (1987). Lifeworld and system: A critique of functionalist reason. (No Title).

Heng, M. S. H., & De Moor, A. (2003). From Habermas's communicative theory to practice on the internet. Information Systems Journal, 13(4), 331–352.

HF Canonical Model Maintainers. (2022). Distilbert-base-uncased-finetuned-sst-2-english. Hugging Face.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

Jigsaw, G. (2017). Perspective API. https://www.perspectiveapi.com/

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G. & others. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. arXiv Preprint arXiv:2205.11916.

Kortemeyer, G. (2023). Could an Artificial-Intelligence agent pass an introductory physics course? arXiv Preprint arXiv:2301.12127.

Leidner, D., & Tona, O. (2021). The CARE Theory of Dignity Amid Personal Data Digitalization. Management Information Systems Quarterly, 45(1).

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv Preprint arXiv:2107.13586.

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., & others. (2023). Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. arXiv Preprint arXiv:2310.19775.

Maree, J. G. (2021). The psychosocial development theory of Erik Erikson: Critical overview. Early Child Development and Care, 191(7–8), 1107–1121.

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. European Journal of Information Systems, 31(3), 257–268.

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: A three-layered approach. AI and Ethics, 1–31.

OpenAI. (2023). GPT-4 Technical Report. arXiv Preprint arXiv:2303.08774.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., & others. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems.

Papalia, D., & Martorell, G. (2023). Experience Human Development (15th ed.). McGraw Hill.

Papalia, D., Olds, S., & Feldman, R. (2008). Human Development. McGraw-Hill Education.

Pérez, J. M., Giudici, J. C., & Luque, F. (2021). pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks.

Porra, J., Lacity, M., & Parks, M. S. (2020). "Can Computer Based Human-Likeness Endanger Humanness?" – A Philosophical and Ethical Perspective on Digital Assistants Expressing Feelings They Can't Have". Information Systems Frontiers, 22(3), 533–547.

Schlagwein, D., Cecez-Kecmanovic, D., & Hanckel, B. (2019). Ethical norms and issues in crowdsourcing practices: A Habermasian analysis. Information Systems Journal, 29(4), 811–837.

Schneider, J., Abraham, R., Meske, C., & Brocke, J. V. (2022). Artificial Intelligence Governance For Businesses. Information Systems Management.

Schneider, J., Meske, C., & Kuss, P. (2024). Foundation Models. Business Information Systems Engineering.

Schneider, J., Richner, R., & Riser, M. (2023). Towards trustworthy autograding of short, multi-lingual, multi-type answers. International Journal of Artificial Intelligence in Education, 33(1), 88–118.

Schneider, J., Schenk, B., Niklaus, C., & Vlachos, M. (2023). Towards LLM-based Autograding for Short Textual Answers. arXiv Preprint arXiv:2309.11508.

Schöbel, S., Schmitt, A., Benner, D., Saqr, M., Janson, A., & Leimeister, J. M. (2023). Charting the Evolution and Future of Conversational Agents: A Research Agenda Along Five Waves and New Frontiers. Information Systems Frontiers.

Stahl, B. C., Doherty, N. F., & Shaw, M. (2012). Information security policies in the UK healthcare sector: A critical evaluation. Information Systems Journal, 22(1), 77–94. https://doi.org/10.1111/j.1365-2575.2011.00378.x

Young, A. G. (2018). Using ICT for social good: Cultural identity restoration through emancipatory pedagogy. Information Systems Journal, 28(2), 340–358.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., & others. (2023). A survey of large language models. arXiv Preprint arXiv:2303.18223.