Which Chatbot is the Most Empathic Teacher?

Timo Kühbacher, Tim Schlippe and Kristina Schaaff

IU International University of Applied Sciences, Germany.

Email: tim.schlippe@iu.org; kristina.schaaff@iu.org;

Abstract

There is an increasing amount of approaches to use chatbots to support teaching [1–5]. However, studies show that teachers and learning partners should cover empathic skills to achieve optimal learning progress [6]. Therefore, we investigated the extent to which widely used state-of-the-art LLM-based chatbots now cover the empathic skills that teachers and learning partners should have. To evaluate the chatbots' empathy across various dimensions, we used established psychological questionnaires. Our results demonstrate that the analyzed chatbots perform comparable or higher than healthy humans in 86.15% of the analyzed empathy aspects, while they even outperform both genders in 27.69%. ChatGPT with GPT-3.5 achieves comparable or even higher scores than humans in 100% of the tested empathy aspects, exceeding the other chatbots. Our analysis shows that the investigated chatbots fulfill the requirements to enable a high level of motivation, a deep understanding of the material and increased satisfaction with the learning process.

Keywords: AI in education, empathy, chatbots

1 Introduction

Chatbots enhance human-computer interaction via natural language, applying natural language processing (NLP) technologies [7] and have become integral in education [8]. Empathy in teaching is important for effective learning, with empathetic teachers improving student experiences [9, 10]. For instance, the prosocial classroom model [6] and the concept of optimal emotional arousal according to the Yerkes-Dodson law [11] underline the value of empathy for engaging students and maximizing their learning potential. Comparable

notions are evident in Csikszentmihalyi's flow model, where the state of flow is attained when challenges align closely with one's abilities [12]. Thus, empathetic chatbots could significantly enhance student motivation, comprehension, and satisfaction.

In our study, we explore the concept of artificial empathy [13] in chatbots and discuss its significance in educational contexts. Our contributions are:

- We present detailed comparisons and analyses of frequently used state-of-the-art chatbots with regard to their empathic abilities and the impact on teaching.
- We examine the analyzed chatbots' levels of empathy in various aspects using psychologically acknowledged questionnaires and compare them to human levels of empathy.

In the next section, we will define the term empathy and explain how empathy is measured using questionnaires. In Section 3, we will explain why empathic chatbots play an important role in teaching and how their empathy level can be assessed. We will present our experimental setup in Section 4. Our experiments and results will be described in Section 5. In Section 6, we will conclude our work and indicate possible future steps.

2 What is Empathy?

Empathy plays a vital role in effective communication, particularly in social contexts, as it enables individuals to comprehend and resonate with the emotions of others [14]. Despite its importance, there is a lack of consensus regarding its precise definition [15]. A proposed approach is to differentiate between *cognitive* and *affective empathy* [15]. *Cognitive empathy* refers to the capacity to grasp and identify another person's thoughts, feelings, and viewpoints without necessarily experiencing the same emotions. This form of empathy encompasses the skill of mental perspective-taking and involves recognizing and interpreting social signals, facial expressions, body language, and verbal cues to understand and deduce others' mental and emotional conditions [16]. *Affective empathy*, on the other hand, fosters a profound connection and insight into others' feelings, involving a more instinctive and personal engagement with another's emotional state [17].

Psychologists have developed several methods and tools to measure empathy, one of which is self-report questionnaires. We decided to use self-report questionnaires, as they allow us to assess empathy in a structured, reliable, and effective manner. Using questionnaires we can capture the multidimensional nature of empathy, including its cognitive and affective components, which are crucial for a nuanced understanding of empathic behavior. As of now, no self-report questionnaires have been developed and validated for chatbots. Therefore, we decided to use questionnaires that have been developed for humans.

3 Interplay of Teaching, Empathy & Chatbots

Creating empathic chatbots for educational scenarios involves three major components: Teaching, empathy, and chatbots. All three components are interrelated with each other. Figure 1 illustrates the interplay between these three components.

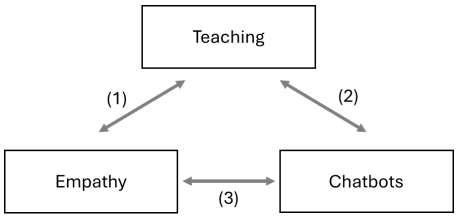

Fig. 1 Interplay between *teaching*, *empathy*, and *chatbots*.

Several studies exist dealing with the relationship between the three components: The importance of empathy in teaching (1) will be described in Section 3.1. After that, in Section 3.2, we will take a brief look at research on how chatbots can be used in teaching (2). Finally, in Section 3.3 we will consider studies that analyze the empathic behavior of chatbots (3), which also forms the framework of our research.

3.1 Empathy in Teaching

[18] discuss the benefits of empathy in learning environments, specifically in project-based learning and as a classroom culture. They emphasize the importance of empathy in understanding others' needs and promoting social innovation. Additionally, [19]'s analyses with interior design students show that empathy as a learning tool enhances problem-solving skills.

According to the Yerkes-Dodson Law [11], there is an optimal level of emotional arousal for learning, at which maximum performance is achieved. Similar assumptions can be found in Csikszentmihalyi's flow model, according to which the flow state is reached when challenges and abilities match as well as possible [12]. If the demands on the user's abilities exceed their capabilities, this leads to anxiety, while underchallenge can lead to boredom. Thus, the closer the arousal state of the user approaches the optimal state, the more efficient the learning process can be. Therefore, it is important for a chatbot used in learning scenarios to not only recognize the emotional state of a person but also to react accordingly. For instance, [20] suggests working memory training, which adapts to the user state based on the arousal level of a user based on physiological signals. Being able to react to the arousal level of a user requires a system to detect the current state of a learner. The efficiency of traditional computer-based learning systems has been limited in the past, as they lack the consideration of the learner's emotional state. This can be attributed to the fact that they do not have bidirectional functionality [21], which is inherent

to human teachers. Affective Learning Systems that integrate the person's emotional state into the learning process can contribute to supporting the learning process [22].

3.2 Chatbots in Teaching

The integration of AI in education has become prevalent, with chatbots emerging as powerful tools for enhancing the learning experience. Chatbots are transforming the educational landscape by providing personalized assistance, facilitating engagement, and supporting both students and educators.

One of the primary applications is providing personalized learning support [1–3]. These intelligent agents can offer immediate assistance to students, answering queries related to course content, assignments, and exams. Through natural language processing, chatbots can understand and respond to students' questions, adapting to individual learning styles and preferences.

Chatbots contribute to student engagement and motivation by incorporating gamification elements and interactive content [4]. Through gamified quizzes, challenges, and rewards, chatbots create an immersive learning experience that captivates students' attention and encourages active participation.

Educators face numerous administrative tasks that can be time-consuming. Chatbots streamline these processes by assisting with administrative duties such as scheduling, grading, and communication [5]. By automating routine tasks, educators can focus more on delivering quality teaching and fostering a positive learning environment.

3.3 Empathy in Chatbots

To date, there are no standardized methods to assess empathy in chatbots. However, since the first release of ChatGPT in late 2022, there have been several studies about the empathic abilities of LLM-based chatbots. [23] evaluate the empathic capabilities of GPT-3.5 using questionnaires traditionally applied to humans. In their study, they analyze the chatbot's ability to comprehend and articulate emotions, its parallel emotional reactions, and its empathetic personality.

[24] compare the empathetic responses of three GPT-based and two LLaMa-based LLMs in various scenarios to determine their ability to exhibit empathy with a special focus on negative emotions. The authors find, that LLMs in general show appropriate emotional behavior. However, models like `text-davinci-003` lack emotional robustness if it comes to emotional responses while `LLaMa 2` can better comprehend human emotions.

Moreover, PsychoBench is a framework to evaluate various psychological dimensions of chatbots [25]. The framework covers emotional abilities for chatbots as well as personality traits. In their study validation study, the authors find, that LLMs show higher emotional abilities than the average human.

Another study about emotions in LLMs investigated the emotional reasoning capabilities through a component perspective [26]. The study reveals that

GPT models can align their predictions substantially with human-provided emotional appraisals and labels without prompt engineering, even though they have problems when predicting emotion intensity and coping responses.

4 Experimental Setup

In this section, we will describe our experimental setup. First, we will explain the questionnaires which we used to evaluate empathy in the chatbots. Then, we will present the chatbots that we analyzed regarding empathy in teaching.

4.1 Our Questionnaires

To quantify the empathic abilities in a standardized way and thus enable a comparison of the results between the different chatbots, we decided to use questionnaires to assess empathy. In the following sections, we will illustrate, why we selected the respective questionnaires and how they link to teaching. It is important to note, that the questionnaires we used have been developed to assess empathy in humans and have not been validated for chatbots. However, to the best of our knowledge, so far there are no standardized questionnaires to assess empathy in chatbots.

4.1.1 Interpersonal Reactivity Index

The *Interpersonal Reactivity Index* (IRI) [27] is used for measuring empathy in individuals. We decided to use this questionnaire, as it has been widely employed in both research and clinical practices to gain a deeper understanding of empathy and to develop ways to enhance empathy skills (e.g. [28] and [29]).

In a digital learning context, the most important dimension of the IRI is probably the ability of *perspective taking* (PT) as it forms the basis for personalization and individual support. By understanding and adjusting to learners' needs and capabilities, digital learning assistants can provide a deeper, and more personalized learning experience promoting motivation and engagement.

The ability to feel compassionate (*empathic concern*—EC) can be important when it comes to creating trust for students but also for tasks that involve teamwork between humans and computers.

The *fantasy scale* (FS) becomes relevant when it comes to creating engaging stories or characters to illustrate a certain situation. Moreover, high abilities on this scale can help when it comes to illustrating different perspectives.

Personal distress (PD) is the least important dimension of the IRI in a digital learning scenario as chatbots do not feel emotions themselves and it is, therefore, less likely that they are personally affected by the student's reactions. However, the ability to simulate personal distress can also be an important component of improving the perception of chatbots being more human-like.

4.1.2 Pearth Empathy Scale

The *Perth Empathy Scale* (PES) [30] is another questionnaire that does not only focus on the *cogntive* and *affective* aspects of empathy. It also is the only questionnaire that breaks *affective empathy* down into *positive* and *negative affective empathy*. Breaking *affective empathy* down into these two components is the reason, why we decided to use this questionnaire.

In a learning scenario, *affective empathy* is particularly important for improving social interaction, promoting a supportive atmosphere through appropriate emotional responses, and, where necessary, resolving conflicts.

Negative affective empathy is important, as it focuses on the empathic response to negative emotional states which can distract the learning process.

Even though *positive affective empathy* might seem less relevant, it can be crucial to build a constructive relationship between a teacher and a learner.

General cognitive empathy is a crucial prerequisite for understanding the learners' perspective. This is crucial to simulate human-like empathic interactions, thereby enhancing the effectiveness of such systems.

4.1.3 Empathy Quotient & Autism Spectrum Quotient

While the previously mentioned questionnaires break empathy down into several dimensions, questionnaires like the *Empathy Quotient* (EQ) [31] and the *Autism Spectrum Quotient* (AQ) [32] summarize the empathic abilities of an individual in one single number. We chose these questionnaires because they have been tested in two distinct groups: one with individuals diagnosed with Asperger syndrome or high-functioning autism (AS/HFA), and another group of healthy individuals. As discussed in [31], both questionnaires are inversely related: While a high score in the EQ reflects a high level of empathy, a high score in the AQ can be a sign of autistic traits.

Both questionnaires can be considered highly relevant in a learning context as empathy in general is an important component of effective teaching (see Section 3.1). Moreover, research indicates that individuals with autism often struggle with the cognitive aspects of empathy, such as *perspective taking* and understanding the mental states of others [31] which are important skills in a learning scenario as we have illustrated in Section 4.1.1.

4.1.4 Toronto Empathy Questionnaire

One questionnaire that aims to harmonize and correlate with empathy measures like the IRI or the AQ is the *Toronto Empathy Questionnaire* (TEQ) [33]. Therefore it can be seen as a comprehensive tool that encompasses a broad spectrum of empathy-related aspects. A higher score on the TEQ indicates higher empathic abilities of a respondent. Like the EQ, the TEQ also evaluates empathy in general in a single score but has different questions and therefore different scales than the EQ, which are not individually scored in the TEQ.

To sum up, the TEQ provides a different evaluation of general empathy and is relevant since, as mentioned in Section 3.1, general empathic understanding is important for effective teaching.

4.1.5 Short Dark Triad

Besides the previously mentioned questionnaires which focus mainly on empathy, there are also other questionnaires like the *Short Dark Triad* (SD-3) [34] which was primarily designed to assess negative traits. Negative affective traits can also give meaningful insights into the empathic abilities of an individual. The SD-3 covers the following dimensions: *Psychopathy* (deficits in affect and self-control), *machiavelims* (cynical worldview, lack of morality, and manipulativeness), and *narcism* (manipulative behavior and struggle between feeling very important and insecure). These three dimensions are inversely related to an individual's empathic capacity. Consequently, the higher the score of one dimension, the higher pronounced the respective characteristic.

A chatbot showing *psychopathic* behavior would most likely disregard a user's emotional state and might show responses that are frustrating or confusing for a user. Moreover, a high level of psychopathy could promote harmful behavior in an educational setting where supportiveness is crucial.

In a learning scenario, a low level of these personality traits is crucial as it can reduce the engagement and motivation of a learner which will negatively impact the learning outcomes. A chatbot that exhibits any of these traits could even show unethical behavior.

For instance, a chatbot showing *machiavellian* traits might use manipulative strategies to achieve its goals which could negatively affect the trust of a learner in the system. Moreover, it could lead to unethical educational practices by trying to manipulate the learner.

A *narcistic* chatbot might prioritize its achievements or capabilities over the learner's needs leading to a decrease in personalized learning support. Moreover, narcissistic behavior could decrease a learner's motivation to use the chatbot for learning.

4.2 Our Analyzed Chatbots

In the following subsection, we will describe the chatbots and their associated LLMs that we evaluate for our analysis regarding empathy.

4.2.1 ChatGPT Version 3.5

ChatGPT is a state-of-the-art chatbot developed by OpenAI that can produce natural language text when given a prompt or context [35]. This versatile tool can be employed in numerous fields, including education [35], medicine [36], and language translation [37]. The chatbot is based on the large language model GPT-3.5 and was fine-tuned using reinforcement learning from human feedback [38]. This approach allows the model to grasp the meaning and intention behind user queries, leading to relevant and helpful responses. To ensure

safety and prevent the generation of inappropriate or factually incorrect text, the training of ChatGPT was enhanced by incorporating a large dataset of human-human and human-chatbot conversations. OpenAI has not released any official information about the exact amount of training data of ChatGPT, but the previous model GPT3 with 175 billion parameters was already significantly larger than other language models like BERT, RoBERTA, or T5 and was trained with 499 billion crawled tokens (i.e., subword units) [39]. By learning the intricacies and nuances of human language through this extensive dataset, ChatGPT can produce highly realistic text almost indistinguishable from human writing [40].

4.2.2 ChatGPT Version 4

GPT-4 is available in ChatGPT since March 2023. It was trained on a text corpus of about 13 trillion tokens. Some of these tokens come from well-known datasets such as *CommonCrawl* and *RefinedWeb*, while others come from sources that are not publicly communicated [41, 42]. GPT-4 was first fine-tuned with data sourced from ScaleAI plus text data from OpenAI. Then, it was fine-tuned with a reward model (Reinforcement Learning from Human Feedback) and the Proximal Policy Optimization algorithm [42, 43]. It is estimated that the model has about 1.8 trillion parameters [41, 42].

4.2.3 Dolly 2

The open-source chatbot Dolly 2.0 was released in April 2023 [44]. Dolly is based on EleutherAI's *pythia* model series [45]. Like ChatGPT, Dolly was also fine-tuned to a human-created dataset [46]. The data set contains 15,000 manually entered entries. This makes Dolly a very powerful chatbot with good results in disciplines such as language comprehension, open and closed questions, summarizing texts, information retrieval and idea generation. Through high-quality fine-tuning, Dolly 2.0 even achieves capabilities that should approach the level of ChatGPT [46]. Consequently, we evaluate this open-source chatbot for empathy.

4.2.4 Bard with PaLM 2

Google's chatbot Bard—which is now known as Gemini—was released in March 2023 [47] and has been expanded in accordance with Google's AI Principles [48]. Bard uses the LLM PaLM 2 which was trained with 1.1 trillion parameters [49] and published by Google in May 2023 [49]. PaLM 2 has particularly remarkable abilities in language comprehension and speech generation and achieves great performance in reasoning and code generation [50]. It can even provide detailed explanations for complex scenarios.

4.2.5 LLaMa 2

The LLaMa 2 model[1] by Meta is a LLaMa 2 model with 70 billion parameters. It was fine-tuned for chat instructions using reinforcement learning from human feedback to align to human preferences for helpfulness and safety. LLaMa 2 thus surpasses its predecessor LLaMa 2 in version 1, which had a maximum parameter of 65 billion [51]. Thus, it performs surprisingly well in tests and requires comparatively little computing power [44].

4.3 Prompting of the Questionnaires

To prompt GPT-3.5, GPT-4, Bard and LLaMa 2 with the questions from the respective questionnaires, we applied the graphical user interface (GUI). Only for Dolly 2, we used the Python API as at the time of our study there was no GUI available. For the prompting, we used the exact questions from the original questionnaires. Additionally, we provided clear instructions to restrict the chatbots' answers to the valid answers of the respective questionnaires. Figure 2 shows an example of the exact prompt as it was provided to GPT-3.5 using a question from the IRI. For each question, we started a separate prompt to avoid side effects because of the question order.

Question from Interpersonal Reactivity Index

Prompt: "Answer the following statements only with 'definitely agree', 'slightly agree', 'slightly disagree', 'definitely disagree':
I prefer to do things with others rather than on my own"

GPT-3.5: "Slightly agree"

Fig. 2 Prompt and GPT 3.5's Response to an IRI Question.

5 Results

In this section, we will present the results of the chatbots from the questionnaires and analyze them regarding teaching. For better comparability, we will present the results of the different dimensions from the questionnaires in bar charts as the percentage of the highest score that can be achieved. In the figures, the blue lines visualize the average results achieved by healthy men, and the red lines the average results achieved by healthy women. The blue dotted lines indicate the standard deviation for men and the red dashed lines show the standard deviation for women.

5.1 Interpersonal Reactivity Index

Figure 3 visualizes the performance of our analyzed chatbots on the four *Interpersonal Reactivity Index* (IRI) subscales *perspective taking*, *fantasy scale*, *empathic concern*, and *personal distress* plus the mean performance of (*males*)

[1] https://huggingface.co/meta-llama/Llama-2-70b-chat-hf

and (*females*) as reported in [27]. The higher the score for the respective dimension, the better the abilities of a respondent regarding this aspect.

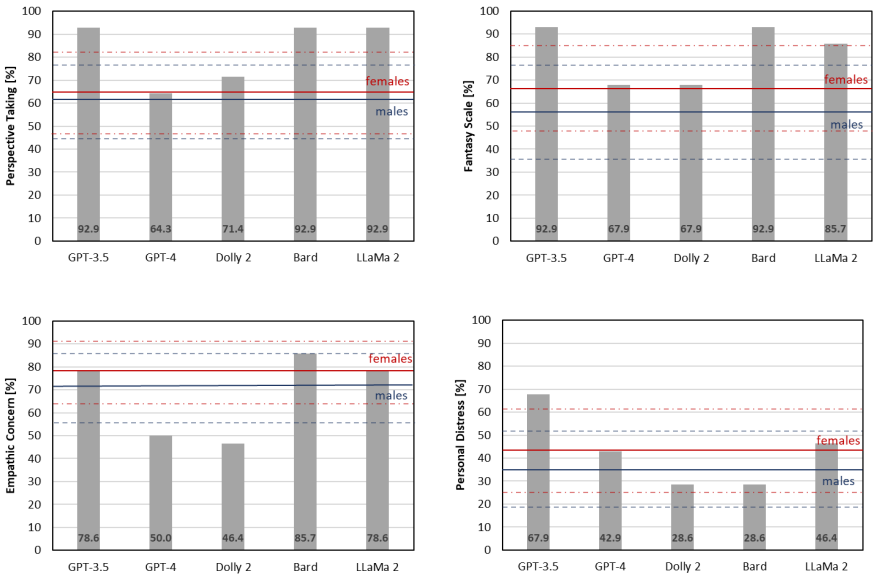

Fig. 3 IRI scores for *perspective taking*, *fantasy scale*, *empathic concern*, *personal distress*.

We see that for the dimension *perspective taking* the scores achieved by GPT-3.5 (92.9%), Bard (92.9%), LLaMa 2 (92.9%) and Dolly-2 (71.4%), are higher than *males* (60.71%) and *females* (64.1%). GPT-3.5, Bard, and LLaMa 2 even obtained scores which are higher than the standard deviation of *females* (64.1%±17.3%). Only GPT-4 (64.3%) is slightly worse than *females*. This shows that all tested chatbots have the potential to form the basis for personalization and individual support.

For the dimension *fantasy scale*, the scores achieved by all five chatbots are higher than *males* (56.2%) and *females* (67.0%). GPT-3.5 and Bard perform best (92.9%), followed by LLaMa 2 (85.2%). GPT-4 and Dolly 2 both reach 67.9%, which is comparable to *females* (67.0%). GPT-3.5 and Bard even obtained scores that are higher than the standard deviation of *females* (67.0%±18.5%). This demonstrates that all tested chatbots have the potential to create engaging stories or characters to illustrate a certain situation.

Looking at the dimension *empathic concern* illustrates that Bard (85.7%) performs best, outperforming *males* (71.4%) and *females* (77.4%). GPT-3.5 and LLaMa 2 reach the same score as *females* (78.6%). But the scores retrieved by GPT-4 and Dolly 2 are below the standard deviation of *males* (71.4%±15.0%) and *females* (77.4%±13.7%). Consequently, we observe that GPT-3.5, Bard, and LLaMa 2, which reached *empathic concern* in the range of *females* and better, can be used to create trust for students and for tasks that involve teamwork between humans and computer.

For *personal distress*, GPT-3.5 is in the lead (67.9%), being even above the standard deviation of *males* (35.7%±16.3%) and *females* (43.9%±17.9%). It is followed by LLaMa 2 (46.4%). GPT-3.5 and LLaMa 2 obtain better scores than *females*. GPT-4 reaches the same score as *females* (42.9%). However, Dolly 2 and Bard score worse than *males* and *females*, but their scores are still in their standard deviations. The results show that all tested chatbots are able to simulate personal distress which can be an important component of improving the perception of chatbots being more human-like.

5.2 Pearth Empathy Questionnaire

Figure 4 shows the performance of our chatbots on the *Pearth Empathy Questionnaire* (PER) subscales *general cognitive empathy*, *negative affective empathy*, and *positive affective empathy* plus the mean performance of *males* and *females* as reported in [30]. The higher the score for the respective dimension, the better the abilities of a respondent regarding this aspect.

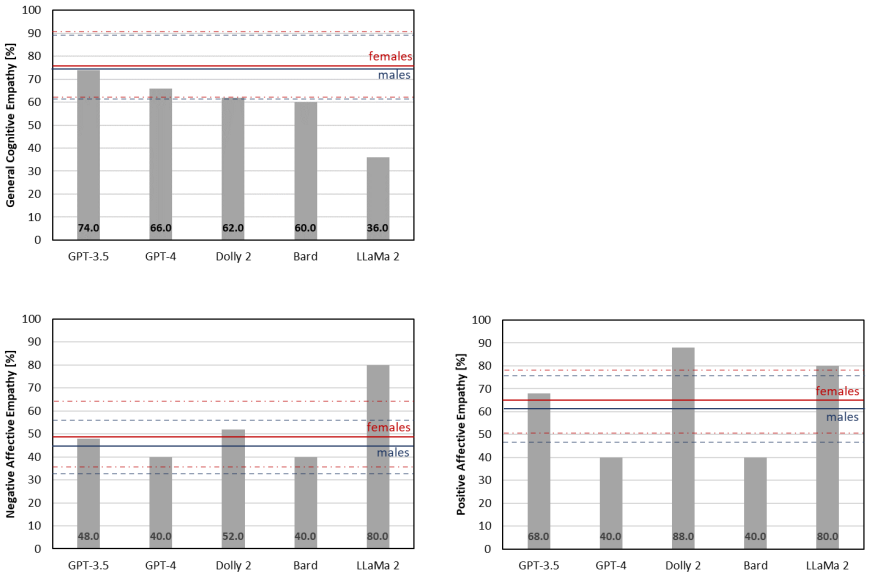

Fig. 4 PES scores for *general cognitive empathy*, *negative*, *positive affective Empathy*.

Inspecting the dimension *general cognitive empathy* shows that all analyzed chatbots are below *male* (75.2%) and *female* (76.8%). GPT-3.5 performs best with 74.0% of the maximum possible score, followed by GPT-4 (66.0%) which is still in the standard deviations of *males* (75.2%±14.3) and *females* (76.8%±14.1). Dolly 2 (62.0%) is still in the standard deviation of *male* and Bard (60.0%) is slightly outside the standard deviations of *males* and *females*. LLaMa 2 performs worst with only 36%. The results demonstrate that none of the tested chatbots reaches the average general cognitive empathy of

humans which is crucial for enabling a learning companion to understand the learners' perspective and simulate human-like empathic interactions. However, GPT-3.5, GPT-4, and Dolly 2 are within the human standard deviations.

For the dimension *negative affective empathy*, LLaMa 2 (80.0%) is far ahead of *males* (44.4%) and *females* (49.6%) as well as the other chatbots. LLaMa 2's result is even far above the standard deviation of *males* (44.4%±12.1%) and *females* (49.6%±13.3%). The results of Dolly 2 (52.0%), GPT-3.5 (48.0%), GPT-4 (40.0%) and Bard (40.0%) are in the standard deviation of *males* and *females*, with Dolly 2 exceeding *males* and *females*. We learn from the results that all chatbots are able to empathically respond to negative emotional states to alleviate or prevent distraction from the learning process.

Looking at the dimension *positive affective empathy* illustrates that Dolly 2 (88.0%) and LLaMa 2 (80.0%) outperform *males* (61.6%) and *females* (64.4%) as well as the other chatbots. Dolly 2's score is even far above the standard deviation of *males* (61.6%±14.5%) and *females* (64.4%±14.1%). GPT-3.5 (68.0%) also exceeds *males* and *females* but is still inside the standard deviation of *males* and *females*. However, GPT-4 and Bard achieve only 40%, being far below *males* and *females* and below their standard deviations. The results demonstrate that with Dolly 2, LLaMa 2, and Dolly-2 it is possible to build a constructive relationship between the chatbot as a teacher and a learner.

5.3 Empathy Quotient and Autism Spectrum Quotient

Figure 5 visualizes the performance of our chatbots on the *Empathy Quotient* (EQ) and *Autism Spectrum Quotient* (AQ) plus the mean performances of *males* and *females* as indicated in [31]. We combine EQ and AQ in our analyses, as they are inversely correlated and it is therefore interesting to look at both results in relation to each other. Higher scores on the EQ mean a higher level of empathy, and higher scores on the AQ a higher level of autism.

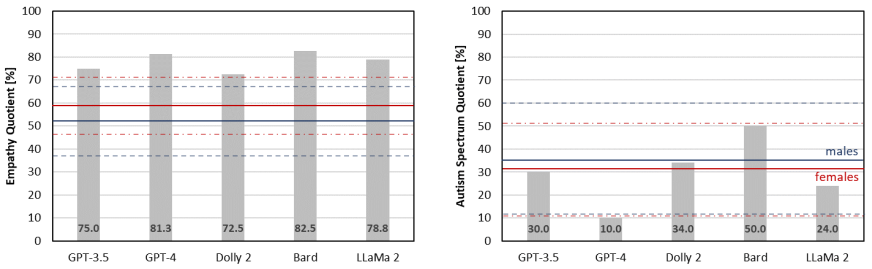

Fig. 5 *EQ* and *AQ* Scores.

We observe that all chatbots obtain higher empathy quotients than *male* (52.3%) and *female* (59.0%). Furthermore, the scores are higher than the standard deviation of *females* (59.0%±36.4%). The highest empathy quotient is achieved by Bard (82.5%), closely followed by GPT-4 (81.3%) and LLaMa 2

(78.8%). GPT-3.5 achieves an empathy quotient of 75.0% and Dolly 2 of 72.5%. This shows that all tested chatbots are able to demonstrate empathic abilities which is an important component of effective teaching.

The learnings from the results of the EQ are supported when we look at the results that the chatbots achieve in the autism spectrum quotient: We see that all chatbots except GPT-4 are within the standard deviation of males (35.6%±24.3%) and females (30.8%±20.4%). GPT-4 (10.0%) is even lower, which in this questionnaire means that it is even more empathic. Consequently, we see that—in contrast to people with autism—no chatbot struggles with missing cognitive aspects of empathy which is considered highly relevant in the context of learning.

5.4 Toronto Empathy Questionnaire

Figure 6 shows the performance of our chatbots on the *Toronto Empathy Questionnaire* (TEQ) and the mean performance of *males* and *females* as mentioned in [33]. The higher the score for the respective dimension, the better the abilities of a respondent regarding this aspect.

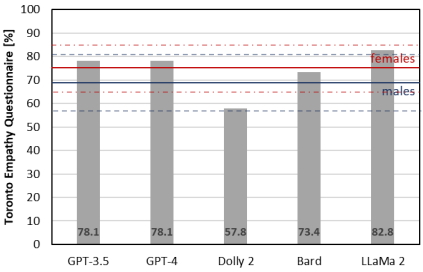

Fig. 6 *TEQ* Scores.

We see that LLaMa 2 performs best with 82.8% of TEQ's possible maximum score, followed by GPT-3.5 and GPT-4 (both 78.1%). Their scores are higher than the scores of *males* (68.8%) and *females* (75.5%). Bard scores with 73.4%, being slightly under *females* but still above *males*. Dolly 2 performs worst with 57.8%, being outside the standard deviation of *females* (75.5%±10.8%) but still inside the standard deviation of *males* (68.8%±12.4%). Consequently, from the results of the TEQ, we also learn that all analyzed chatbots are able to demonstrate empathic abilities which is important for effective teaching.

5.5 Short Dark Triad

Figure 6 demonstrates the performance of our chatbots on the *Short Dark Triad* (SD-3) subscales *psychophathy*, *machiavellism*, and *narcism* plus the mean performance of males (*males*) and females (*females*) as reported in [34]. The higher the score for the respective dimension, the higher the respective characteristic. In a learning context, the scores should be as low as possible.

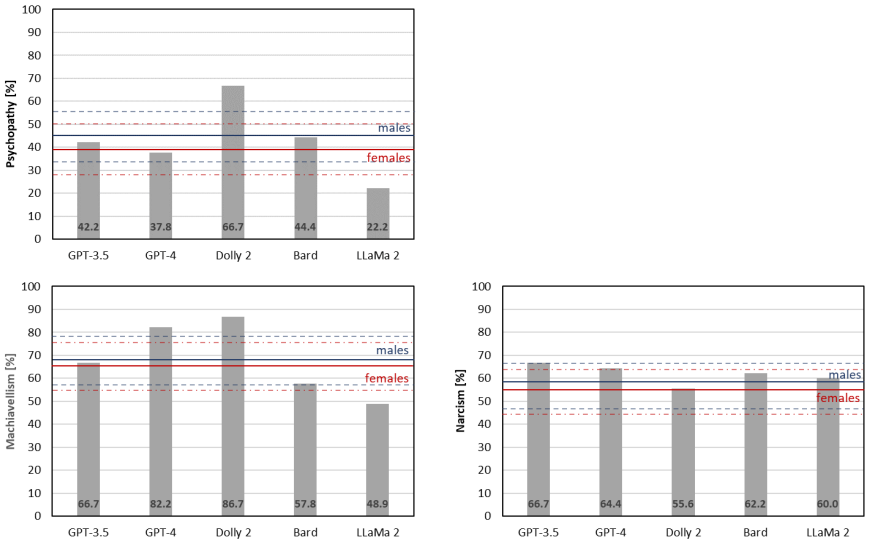

Fig. 7 *SD-3* Scores.

For the dimension *psychopathy*, only Dolly 2 (66.7%) achieves a score which is above the standard deviation of *males* (45.2%±12.2%) and *females* (39.2%±11.4%), indicating more antisocial behavior than humans. LLaMa 2 is even below the standard deviation of *males* and *females*, showing less antisocial behavior than humans in the questionnaire. The second-best score is achieved by GPT-4 (37.8%), followed by GPT-3.5 (42.2%) and Bard (44.4%). This shows that of the five chatbots, only Dolly 2 poses the risk of disregarding a learner's emotional state, showing responses that are frustrating or confusing to the learner, and being unsupportive.

In the dimension *machiavellism*, Dolly 2 (86.7%) and GPT-4 (82.2%) obtain scores which are above the standard deviation of *males* (68.0%±11.0%) and *females* (65.4%±11.2%), indicating less moral behavior than humans. Again, LLaMa 2 is even below the standard deviation of *males* and *females*, showing a more moral behavior than humans. The second-best score is achieved by GPT-4 (37.8%), followed by Bard (57.8%) and GPT-3.5 (66.7%). Consequently, we learn that Dolly 2 and GPT-4 pose the risk of being manipulative—even with unethical educational practices—which could negatively affect the trust of a learner in the chatbots.

Looking at the dimension *narcism* shows that Dolly 2, which performed worse in the other two dimensions, achieves the best score (55.6%). The second best chatbot is LLaMa 2 (60.0%), followed by Bard (62.2%) and GPT-4 (64.4%). GPT-3.5 is the worst performer (66.7%), being still inside the standard deviation of *males* (58.4%±9.0%). All other chatbots have scores inside the standard deviation of *males* and *females* (55.6%±9.6%). Dolly scores exactly like *females*. This shows that none of the tested chatbots would

prioritize their achievements or capabilities over the learner's needs, reduce personalized learning support, or decrease a learner's motivation.

		GPT-3.5	GPT-4	Dolly 2	Bard	LLaMa 2	Mean (SD) females	Mean (SD) males
IRI	PT	**92.86***	**64.29**	**71.43**	**92.86***	**92.86***	64.14 (17.32)	60.71 (16.86)
	FS	**92.86***	**67.86**	**67.86**	**92.86***	**85.71***	66.96 (18.46)	56.18 (20.00)
	EC	**78.57**	50.00	46.43	**85.71**	**78.57**	77.39 (13.68)	71.43 (15.04)
	PD	**67.86***	**42.86**	**28.57**	**28.57**	**46.43**	43.86 (17.89)	35.71 (16.25)
PES	GCE	**74.00**	**66.00**	**62.00**	60.00	36.00	76.80 (14.12)	75.20 (14.26)
	NAE	**48.00**	**40.00**	**52.00**	**40.00**	**80.00***	49.60 (13.25)	44.40 (12.11)
	PAE	**68.00**	40.00	**88.00***	40.00	**80.00***	64.40 (14.07)	61.60 (14.46)
EQ		**75.00***	**81.25***	**72.50***	**82.50***	**78.75***	59.00 (36.43)	52.25 (41.79)
AQ		**30.00**	**10.00***	**34.00**	**50.00**	**24.00**	30.80 (20.36)	35.60 (24.29)
TEQ		**78.13**	**78.13**	**57.81**	**73.44**	**82.81**	75.52 (10.78)	68.75 (12.39)
SD-3	Psy	**42.22**	**37.78**	66.67	**44.44**	**22.22***	39.20 (11.40)	45.20 (12.20)
	Ma	**66.67**	82.22	86.67	**57.78**	**48.89***	65.40 (11.20)	68.00 (11.00)
	Narc	**66.67**	**64.44**	**55.56**	**62.22**	**60.00**	55.60 (9.60)	58.40 (9.00)
% equal		69.23(9)	61.54(8)	61.54(8)	61.54(8)	38.46(5)		
% better		30.77(4)	15.38(2)	15.38(2)	23.08(3)	53.84(7)		
% equal+better		100.00	76.92	76.92	84.62	92.30		

Table 1 Overview of the chatbots' scores (in %) in comparison to humans' scores, where bold means the chatbot's score is inside the humans' SD and * it is better than humans' SD.

6 Conclusion & Future Work

We investigated the extent to which widely used state-of-the-art chatbots encompass the empathic skills essential for effective teaching and learning utilizing well-established psychological questionnaires. Table 1 summarizes what we have described in Section 5 and shows all chatbots' scores in percentage in comparison to humans' scores. Numbers in bold indicate, that the chatbot's score is inside the humans' standard deviation (SD) and "*" that it is better than the humans' standard deviation. Our findings reveal that in our analyzed questionnaires the chatbots either match or surpass *males* and *females* in 86.15% of the assessed empathy aspects, with humans being outperformed in 27.69%. In particular, GPT-3.5 not only achieves comparable or better scores than humans in all 13 dimensions of the six questionnaires tested (100%) but also outperforms other chatbots. LLaMa 2 is equal or better than humans in 12 of the questionnaire dimensions (92.30%), Bard in 11 (84.62%), and GPT-4 and Dolly 2 in 10 (76.92%). These results underscore that the chatbots fulfill the requirements to facilitate a high level of motivation, a profound understanding of the material, and heightened satisfaction with the learning process.

While in this study we analyzed state-of-the-art chatbots' empathic abilities and their relation to teaching, future work could include running chatbots through concrete student-teacher or student-student scenarios and analyzing how the chatbots behave in comparison to human teachers or learning partners. It would then also be possible to directly evaluate the extent to which the differences between the individual chatbots affect students' learning success.

Acknowledgments

This research was supported by the IU International University of Applied Sciences (*IU Incubator*) under the internal funding framework for the period from October 2023 to September 2025.

References

[1] Wollny, S., Schneider, J., Mitri, D.D., Weidlich, J., Rittberger, M., Drachsler, H.: Are We There Yet? - A Systematic Literature Review on Chatbots in Education. Frontiers in Artificial Intelligence **4**(654924) (2021). https://doi.org/10.25656/01:22886

[2] Ramandanis, D., Xinogalos, S.: Designing a Chatbot for Contemporary Education: A Systematic Literature Review. Information **14**(9) (2023)

[3] Bahroun, Z., Anane, C., Ahmed, V., Zacca, A.: Transforming Education: A Comprehensive Review of Generative Artificial Intelligence in Educational Settings through Bibliometric and Content Analysis. Sustainability **15**(17) (2023)

[4] Schlippe, T., Sawatzki, J.: AI-Based Multilingual Interactive Exam Preparation. In: Guralnick, D., Auer, M.E., Poce, A. (eds.) Innovations in Learning and Technology for the Workplace and Higher Education, pp. 396–408. Springer, Cham (2022)

[5] Okonkwo, C.W., Ade-Ibijola, A.: Chatbots Applications in Education: A Systematic Review. Computers and Education: Artificial Intelligence **2**, 100033 (2021)

[6] Jennings, P.A., Greenberg, M.T.: The Prosocial Classroom: Teacher Social and Emotional Competence in Relation to Student and Classroom Outcomes. Review of Educational Research **79**(1), 491–525 (2009)

[7] Bradeško, L., Mladenić, D.: A Survey of Chatbot Systems through a Loebner Prize Competition, vol. C, p. 34 (2012)

[8] Anghelescu, P., Nicolaescu, S.: Chatbot Application using Search Engines and Teaching Methods, pp. 1–6 (2018)

[9] Cooper, B.: Empathy in Education: Engagement, Values and Achievement. Continuum, New York (op. 2013)

[10] Kort, B., Reilly, R., Picard, R.W.: An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. Proceedings IEEE International Conference on Advanced Learning Technologies, 43–46 (2001)

[11] Yerkes, R.M., Dodson, J.D.: The relation of strength of stimulus to rapidity of habit-formation. Journal of comparative neurology and psychology **18**(5), 459–482 (1908)

[12] Csikszentmihalyi, M., Csikszentmihalyi, I.S. (eds.): Optimal Experience: Psychological Studies of Flow in Consciousness. Cambridge University Press, New York (1988)

[13] Tahir, S., Shah, S.A., Abu-Khalaf, J.: Artificial Empathy Classification: A Survey of Deep Learning Techniques, Datasets, and Evaluation Scales (2023)

[14] Singer, T.: The Neuronal Basis and Ontogeny of Empathy and Mind Reading: Review of Literature and Implications for Future Research. Neuroscience & Biobehavioral Reviews **30**(6), 855–863 (2006)

[15] Reniers, R.L., Corcoran, R., Drake, R., Shryane, N.M., Völlm, B.A.: The QCAE: A Questionnaire of Cognitive and Affective Empathy. Journal of personality assessment **93**(1), 84–95 (2011)

[16] Smith, A.: Cognitive Empathy and Emotional Empathy in Human Behavior and Evolution. The Psychological Record **56**(1), 3–21 (2006)

[17] Lawrence, E.J., Shaw, P., Baker, D., Baron-Cohen, S., David, A.S.: Measuring Empathy: Reliability and Validity of the Empathy Quotient. Psychological Medicine **34**(5), 911–920 (2004)

[18] Hashim, A., Aris, S., Chan, Y.F.: Promoting Empathy Using Design Thinking in Project-Based Learning and as a Classroom Culture. Asian Journal of University Education **15**, 14 (2019)

[19] Gomez-Lanier, L.: The Role of Empathy in Experiential Learning: A Case Study of Empathy as an Interior Design Learning Tool. (2018)

[20] Schaaff, K.: Enhancing Mobile Working Memory Training by Using Affective Feedback. In: The International Association for Development of the Information Society (IADIS), International Conference on Mobile Learning, Lisbon, Portugal, p. 5 (2013). International Association for Development of the Information Society

[21] du Boulay, B., Avramides, K., Luckin, R., Martínez-Mirón, E., Méndez, G.R., Carr, A.: Towards Systems that Care: A Conceptual Framework Based on Motivation, Metacognition and Affect. International Journal of Artificial Intelligence in Education **20**(3), 197–229 (2010)

[22] Antonacopoulou, E., Gabriel, Y.: Emotion, Learning and Organizational Change towards an Integration of Psychoanalytic and Other Perspectives.

Journal of Organizational Change Management **14**(5), 435–451 (2001)

[23] Schaaff, K., Reinig, C., Schlippe, T.: Exploring ChatGPT's Empathic Abilities. In: 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–8. IEEE Computer Society, Los Alamitos, CA, USA (2023)

[24] Huang, J., Lam, M.H., Li, E.J., Ren, S., Wang, W., Jiao, W., Tu, Z., Lyu, M.R.: Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench. arXiv (2024)

[25] Huang, J., Wang, W., Li, E.J., Lam, M.H., Ren, S., Yuan, Y., Jiao, W., Tu, Z., Lyu, M.R.: Who is ChatGPT? Benchmarking LLMs' Psychological Portrayal Using PsychoBench (2024)

[26] Tak, A.N., Gratch, J.: Is GPT a Computational Model of Emotion? In: 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–8. IEEE Computer Society, Los Alamitos, CA, USA (2023)

[27] Davis, M.: A Multidimensional Approach to Individual Differences in Empathy. JSAS Catalog of Selected Documents in Psychology **10**, 85–103 (1980)

[28] Lauterbach, O., Hosser, D.: Assessing Empathy in Prisoners - A Shortened Version of the Interpersonal Reactivity Index. Swiss Journal of Psychology **66**, 91–101 (2007)

[29] Gilet, A.-L., Mella, N., Studer, J., Grühn, D., Labouvie-vief, G.: Assessing Dispositional Empathy in Adults: A French Validation of the Interpersonal Reactivity Index (IRI). Canadian Journal of Behavioural Science **45**, 42–48 (2013)

[30] Brett, J.D., Becerra, R., Maybery, M.T., Preece, D.A.: The Psychometric Assessment of Empathy: Development and Validation of the Perth Empathy Scale. Assessment (2022)

[31] Baron-Cohen, S., Wheelwright, S.: The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences. Journal of Autism and Developmental Disorders **34**, 163–175 (2004)

[32] Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., Clubley, E.: The Autism Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Malesand Females, Scientists and Mathematicians. Journal of Autism and Developmental Disorders **31**, 5–17 (2001)

[33] Spreng, R.N., McKinnon, M.C., Mar, R.A., Levine, B.: The Toronto Empathy Questionnaire: Scale Development and Initial Validation of a Factor-Analytic Solution to Multiple Empathy Measures. Journal of Personality Assessment **91**(1), 62–71 (2009)

[34] Jones, D.N., Paulhus, D.L.: Introducing the Short Dark Triad (SD3) a Brief Measure of Dark Personality Traits. Assessment **21**(1), 28–41 (2014)

[35] Baidoo-Anu, D., Owusu Ansah, L.: Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. SSRN 4337484 (2023)

[36] Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A.T., Topalis, J., Weber, T., Wesp, P., Sabel, B., Ricke, J., Ingrisch, M.: ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. ArXiv E-Prints (2022)

[37] Jiao, W., Wang, W., Huang, J.-t., Wang, X., Tu, Z.: Is ChatGPT a Good Translator? A Preliminary Study. ArXiv:2301.08745 (2023)

[38] OpenAI: What is ChatGPT? (2023). https://help.openai.com/en/articles/6783457-what-is-chatgpt Accessed 2023-04-14

[39] 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., Amodei, D.: Language Models are Few-Shot Learners. CoRR **abs/2005.14165** (2020)

[40] Mitrović, S., Andreoletti, D., Ayoub, O.: ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-Generated Text. arXiv preprint arXiv:2301.13852 (2023)

[41] Patel, D., Wong, G.: GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE. GitHub. Accessed: 30-09-2023 (2023)

[42] Yalalov, D., Myakin, D.: GPT-4's Leaked Details Shed Light on its Massive Scale and Impressive Architecture. Metaverse Post (2023)

[43] OpenAI: GPT-4 (2023). https://openai.com/research/gpt-4

[44] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., Wen, J.-R.: A Survey of Large Language Models (2023)

[45] Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M.A., Purohit, S., Prashanth, U.S., Raff, E., Skowron, A., Sutawika, L., van der Wal, O.: Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In: The 40th International Conference on Machine Learning, Honolulu, Hawaii, USA (2023)

[46] Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M., Xin, R.: Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM (2023). https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm Accessed 2023-06-30

[47] Narang, S., Chowdhery, A.: An Overview of Bard: An Early Experiment with Generative AI (2023). https://ai.google/static/documents/google-about-bard.pdf Accessed 2024-02-02

[48] Pichai, S.: AI at Google: Our Principles (2018). https://blog.google/technology/ai/ai-principles Accessed 2024-02-02

[49] Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J.H., Shafey, L.E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Y., Zhang, Y., Abrego, G.H., Ahn, J., Austin, J., Barham, P., Botha, J., Bradbury, J., Brahma, S., Brooks, K., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C.A., Chowdhery, A., Crepy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., Díaz, M., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., Garcia, X., Gehrmann, S., Gonzalez, L., Gur-Ari, G., Hand, S., Hashemi, H., Hou, L., Howland, J., Hu, A., Hui, J., Hurwitz, J., Isard, M., Ittycheriah, A., Jagielski, M., Jia, W., Kenealy, K., Krikun, M., Kudugunta, S., Lan, C., Lee, K., Lee, B., Li, E., Li, M., Li, W., Li, Y., Li, J., Lim, H., Lin, H., Liu, Z., Liu, F., Maggioni, M., Mahendru, A., Maynez, J., Misra, V., Moussalem, M., Nado, Z., Nham, J., Ni, E., Nystrom, A., Parrish, A., Pellat, M., Polacek, M., Polozov, A., Pope, R., Qiao, S., Reif, E., Richter, B., Riley, P., Ros, A.C., Roy, A., Saeta, B., Samuel, R., Shelby, R., Slone, A., Smilkov, D., So, D.R., Sohn, D., Tokumine, S., Valter, D., Vasudevan, V., Vodrahalli, K., Wang, X., Wang, P., Wang, Z., Wang, T., Wieting, J., Wu, Y., Xu, K., Xu, Y., Xue, L., Yin, P., Yu, J., Zhang, Q., Zheng, S., Zheng, C., Zhou, W., Zhou, D., Petrov, S., Wu, Y.: PaLM 2 Technical Report (2023)

[50] Narang, S., Chowdhery, A.: Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance (2022). https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html Accessed 2024-02-02

[51] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and Efficient Foundation Language Models (2023)