UNIVERSITY OF MESSINA

Department of Cognitive, Psychological, Pedagogical and Cultural Studies

PhD in Cognitive Sciences, 38th Cycle

Language Models in Education

Doctoral Thesis by:

Rossella Suriano

Supervisor:

Chiar.mo Prof. Alessio Plebe

PhD Program Coordinator:

Chiar.ma Prof.ssa Alessandra Falzone

SSD: M-FIL/02

A.A. 2022/2023

To those who made this journey possible and wonderful.

Contents

List of Figures

List of Tables

Introduction

The research presented in this thesis aims to contribute to the scientific understanding and analysis of the impact produced by neural language models—one of the most disruptive technological innovations of recent centuries, and one that is poised to increasingly shape human civilization. Within the broader context of the effects of this new AI on human activities, this research has chosen to focus primarily on education, though it is not limited exclusively to this field. It has become evident to all that the rapid spread of AI-based chatbots has triggered profound transformations in the interaction between humans and machines. These sophisticated conversational systems demonstrate an increasing ability to understand human language and provide immediate and personalized responses, increasingly adapting to the needs of users. A pivotal turning point occurred in November 2022, when OpenAI released ChatGPT. From that moment on, the legendary Turing Test (Turing, 1950)—in which a computer engages in a friendly conversation with humans, once deemed unattainable—became a daily reality for millions of people (Perconti & Plebe, 2025).

ChatGPT uses natural language processing to interact in a surprisingly realistic manner. It not only responds to follow-up questions but also admits mistakes, challenges incorrect assumptions, and refuses inappropriate requests, demonstrating a sophisticated understanding of conversation context and nuances (OpenAI, 2023). This wide range of abilities allows it to adapt to different tasks and generate original content like poetry or compelling articles, opening the doors to various fields of application.

In the educational context, the adoption of ChatGPT has raised concerns and questions regarding its responsible use and the impact on students' learning process. The immediate access to seemingly comprehensive answers may lead users to believe that ChatGPT possesses all the necessary information to successfully complete any task. It is crucial to clarify that ChatGPT is not an infallible source of knowledge. Being not an encyclopedia, it might provide incorrect information or seemingly correct answers that are actually incorrect. The conversational and convincing nature of ChatGPT can generate blind trust in its responses, reducing the motivation to thoroughly scrutinize the output. Users may forego verifying or confirming information through multiple reliable sources. By relying solely on the technological system and adopting a passive attitude, the ability to think critically may be limited. Critical thinking allows for evaluating and analyzing the reliability of information before drawing conclusions or making judgments. This type of thinking is transdisciplinary and is described as a complex mental process that involves analyzing information, assessing the reliability of sources, considering different ideas, and making inferences.

Adopting a critical approach in interacting with ChatGPT is essential, but these concerns are not insurmountable, and the adoption of AI-based chatbots in education can be managed responsibly. Moreover, the conversational aspect of ChatGPT can be seen as an advantage in the educational field. Strengths include the ability to create more interactive and engaging learning experiences, allowing students to ask questions and receive immediate answers. Unlike teachers, ChatGPT is always available and can handle multiple requests simultaneously. The use of these systems could have a positive impact on learning by personalizing it and contributing to increased student engagement and motivation. Even concerning critical thinking, ChatGPT could enable users to explore different perspectives and study topics from various points of view. Clearly, it is important to implement functional approaches so that ChatGPT can support the evaluation of users' reasoning processes. In conclusion, the proliferation of ChatGPT is bringing about significant transformations that also impact the educational field. The initial skepticism due to the foreseeable risks should be seen as a starting point to become aware of the limitations inherent in these systems and harness their enormous potential.

In light of the scenario just described, it is clear that the objectives set out by the present research are both urgent and pressing, especially in the field of education and with respect to critical thinking. The first chapter focuses on the technical description of neural language models, illustrating their main underlying mechanisms, such as deep learning and natural language processing. A historical overview of the evolution of chatbots will also be provided, with particular attention to the transition from rule-based systems to generative models. Finally, the chapter will examine the use of chatbots in educational settings, highlighting emerging pedagogical approaches and key research directions in the field. The second chapter is entirely dedicated to the concept of critical thinking. Following a review of the most influential theoretical definitions in the literature, the main cognitive and metacognitive components that characterize this skill will be identified. A standardized tool for measuring critical thinking will then be introduced and described, with the aim of making it applicable both to the assessment of human performance and to the evaluation of output produced by language models. The third chapter presents an experimental study conducted within the framework of machine psychology, an emerging field that aims to investigate patterns of behavior and reasoning in intelligent systems. The study focuses on the ability of neural language models to demonstrate forms of critical thinking through specifically designed tasks, offering both qualitative and quantitative assessments of their performance. The fourth chapter addresses the interaction between students and chatbots, with a particular focus on the potential of tools such as ChatGPT to foster the development of critical thinking in human users. Empirical data collected from real-world educational contexts will be presented and analyzed, with the aim of understanding whether and how such interactions promote reflective, argumentative, and evaluative processes in students. Finally, the fifth chapter broadens the perspective beyond the educational context by exploring the phenomenon of emotional attachment to AI-based chatbots.

Chapter 1

The Paradigm Shift: Language Models Transforming Education

This first chapter aims to provide a comprehensive introduction to the topic of Neural Language Models in education. It begins with their definition, recalls the AI research tradition within which they are historically situated, includes a brief overview of their computational foundations, and finally lists some of their earliest educational applications.

1.1 What Language Models are, and where they come from

This section introduces the concept of Neural Language Models, providing an overview of their origins, development, and key characteristics. A brief terminological note is appropriate here. The term Neural Language Model (NLM) is here preferred, as it explicitly highlights that these AI artifacts belong to the broader family of Artificial Neural Networks. However, it should be made clear from the outset that NLM is essentially synonymous with the now popular term "Large Language Model" (LLM). From a strictly philological standpoint, it is interesting to note that the term LLM was actually introduced well before the invention of the Transformer architecture (Vaswani et al., 2017), which, as will be discussed later, underpins modern NLMs. LLM was originally coined by a Google team in reference to earlier, non-neural n-gram language models (Brants et al., 2007). In the course of this thesis, unless otherwise specified, we will refer to NLMs simply as "Language Models", as is also done in the thesis title. There is no risk of ambiguity, since Language Models earlier than the Transformer - and thus now largely obsolete - are not considered here.

As clear from their name, Language Models are computational models able to process natural language, and they belong to the sector of Artificial Intelligence (AI) known as *Deep Learning* (DL). I will provide a short and sketchy account of DL, restricted to the scope of this thesis, before doing so I provide a few pointers to the existing literature on DL. It is difficult to find a better introduction to the subject than that by the founders of DL themselves (LeCun et al., 2015). Historical accounts of the development of DL can be found in Schmidhuber (2015) and in (Goodfellow et al., 2016, ch.1). A philosophical perspective on DL is provided by Buckner (2018, 2019), a cognitive science perspective can be found in (G. E. Hinton, 2015; Perconti & Plebe, 2020), and accounts of DL in relation with AI in general are in (Boden, 2016; Gerrish, 2018; Pavone & Plebe, 2021; Sejnowski, 2018).

1.1.1 Artificial Neural Networks

DL is a direct descendant of Artificial Neural Networks (ANN), an AI approach loosely inspired by brain neurons. The core idea in ANN and in DL is in the label "learning": these systems are designed to learn all what they should do by experience, without any predefined knowledge. This paradigm is grounded in the philosophical tradition of empiricism. For John Locke (1690) and David Hume (1739) concepts are products of experience, and reason is founded in perception. In the brain neural circuits learn by neural plasticity, the gradual adaptation of several physiological components, like synapses, in response to signal experience (Huttenlocher, 2002; Steeves & Harris, 2013).

The ANN idea was around as early as in the '50s (Minsky, 1954; Rosenblatt, 1958), but it found its way late in the '80s (Rumelhart & McClelland, 1986b). Most ANN are organized as *feedforward networks*, with rows of simple neural elements called layers, and the signals flow from layers to layers only, and not across neurons of the same layer. A sketch of feedforward network is in Fig. 1.1. The connections between each layer and the next one are often called "weights", are the primary source of plasticity that allow the network to learn. The success of the early ANNs was due to an efficient mathematical technique to learn the weights, called *backpropagation*.

All weights between layers are gradually updated using known samples of the function to learn. Being $\vec{w}$ the vector of all weights in a network, and

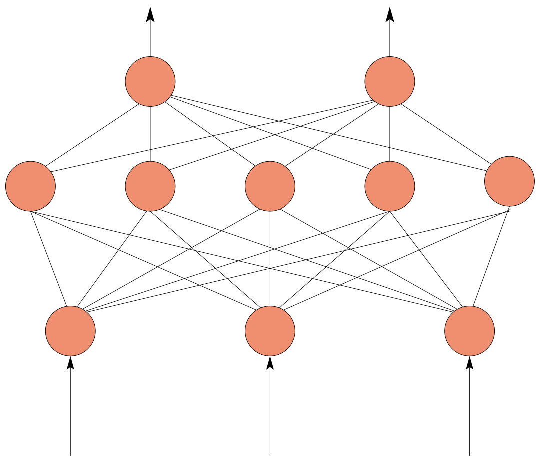


Figure 1.1: Sketch of an artificial neural feedforward network.i Neurons in the input layer get a multidimensional input vector, and neurons in the output layer provide the desired result of the network. The layer in between is called “hidden” because the features represented by each neuron are not explicit, and this layer is responsible for the powerful capabilities of ANNs.

$\mathcal{L}(\vec{x}, \vec{w})$ a measure of the error of the network with weights $\vec{w}$ when applied to the sample $\vec{x}$, the backpropagation updates the parameters iteratively, according to the following formula:

$$\vec{w}_{t+1} = \vec{w}_t - \eta \nabla_w \mathcal{L}\left(\vec{x}_t, \vec{w}_t\right) \tag{1.1}$$

where t spans over all available samples $\vec{x}_t$, and η is the *learning rate*.

Backpropagation was the key factor for a successful season for ANN, however, it worked well only with networks with three layers, like the one in Fig. 1.1. The shift from classical ANN to DL is exactly in overcoming the three layers limitation, and “deep” simply means with many layers. Geoffrey Hinton, who already contributed in inventing the backpropagation (Rumelhart et al., 1986) was the first to train models with four and five hidden layers, using a quite clever combination of training methods (G. E. Hinton & Salakhutdinov, 2006). The three levels wall was broken, and the new wave of research in ANN, renamed DL, boosted the resurgence of AI in the last decade. A new variant of backpropagation is now well established, under the name of *stochastic gradient descent*, is works as following:

$$\vec{w}_{t+1} = \vec{w}_t - \eta \nabla_w \frac{1}{M} \sum_i^M \mathcal{L}\left(\vec{x}_i, \vec{w}_t\right) \tag{1.2}$$

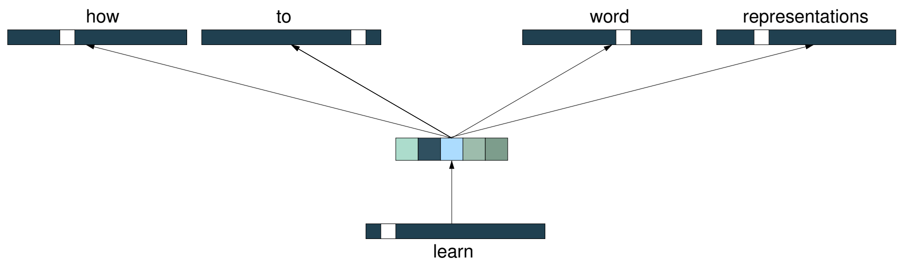


Figure 1.2: Sketch of the word embedding learning by skip-gram. Words are initially coded as simple one-hot vectors. The central word on a sentence is presented in the input layer, the hidden layer is trained to guess a number of previous and next words in the sentence. In this example the sample sentence is `how to learn word representations` and the central word under processing is `learn`.

Instead of computing the gradients over a single sample t as in the original backpropagation, a stochastic estimation is made over a random subset of size M of the entire dataset, and at each iteration step t a different subset, with the same size, is sampled. Even if equations (1.1) and (1.2) are very similar, stochastic gradient descent is able to train models with up to hundreds of layers, while backpropagation had difficulty already with four layers.

1.1.2 Natural Language and Neural Networks

DL suddenly came into view in the landscape of AI, approaching human performances in several domains, like in artificial vision (Krizhevsky et al., 2012), and in defeating the the world champion of *Go*, the most difficult chessboard game in the world (Silver et al., 2016). The progress of DL in the field of natural language processing has been slower, and until 5-6 years ago the performances of DL models remained well under those of a human speaker. There are several reasons that make natural language particularly challenging for ANN.

One obstacle is the clash between the symbolic nature of text and the numerical values of neural signals. The pioneers of ANN in language had to invent a convoluted and implausible solution to code words as vectors of numbers (Rumelhart & McClelland, 1986a), since then researchers in ANN have struggled to find better coding schemes.

A fundamental breakthrough was obtained by Mikolov et al. (2013) of the Google team, with the introduction of the *word embedding* technique.

The main idea is the so-called skip-gram scheme, where a single word inside a sentence is presented to a neural network, which task is to guess the surrounding words in the sentence. In order to run the network a first coding should be established, and is the simple "one-hot vector" solution: vectors have the same dimension of the vocabulary with all values 0 except one, distinct for every word. Of course one-hot vectors have dimensions of hundreds of thousands, while the new learnt vectorial representation is instead very compact, and is made of full valued real vectors. The Fig. 1.2 exemplifies the training procedure for the sample sentence `how to learn word representations`, when the word processed is `learn`, therefore the words to be generated by the model are `how to ...word representations`. A meaningful representation of `learn` will be learned after processing billions of sentenced containing the word `learn`. By "meaningful" we mean that the numerical vectors can be manipulated with results that surprisingly respect lexical semantics aspects. Let $\vec{w}(\cdot)$ being the word embedding transformation, by computing:

$$\vec{q} = \vec{w}\,(\texttt{king}) - \vec{w}\,(\texttt{male}) + \vec{w}\,(\texttt{female})$$

the resulting vector $\vec{q}$ is more similar to $\vec{w}\,(\texttt{queen})$ than to any other word embedded vector. Note how the scheme of Fig 1.2 resembles Hinton's idea of autoencoder, with an encoding layer followed by a decoding layer, in this case however the decoding task is not the same word, but the surrounding words.

A second obstacle in adapting ANNs for language is in the way backpropagation works. Its efficiency come at a price. Neural models can be trained only if there are tasks where precise input and outputs can be specified, and a large set of examples with known correct outputs are available. Language competence is a typical case where there is no clear single task that characterizes its learning. Geoffrey Hinton invented a strategy called *autoencoder* for all cases where a supervised task cannot be easily conceived. The task of the autoencoder is dismaying simple: to produce in output the same data given in input (G. Hinton & Zemel, 1994; G. E. Hinton & Salakhutdinov, 2006). An example of autoencoder is shown in Fig. 1.3. The autoencoder is composed by two distinct submodels: the *encoder* that gets the input values and transform them with severl neural layers, up to an internal representation. This representation is taken as input by the second submodel, the *decoder*, with its own neural layers up to the output, of the same format as the input.

In the context of learning a language the autoencoder task consists in feeding the model with a sentence, and expecting exactly the same sentence

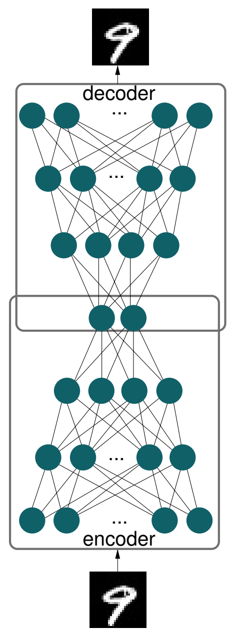


Figure 1.3: Scheme of the autoencoder, in this example applied to the recognition of handwritten digits, in the model by G. E. Hinton and Salakhutdinov (2006).

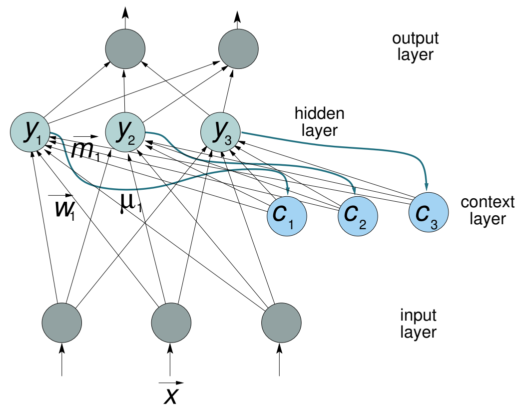


Figure 1.4: Scheme of Elman's neural recurrent network. Note the addition of the "context" layer, parallel to the usual hidden layer.

as output. While engqaged in this simple task, the model builds and internal representation of all what is about in the words of sentences. This incredibly simple task is exactly how all current NLMs are trained.

The last and probably most serious issue with language was the fact that neural networks are static, while language is dynamic. In a classic feedforward network all neural activation depend just on the current input, there is no memory of previous inputs. In a text the meaning of each word is composed in relation with many preceeding words. An important solution was proposed by the psycholinguist Jeffrey Elman (1990) with the so-called *recurrent neural networks*. As shown in Fig. 1.4, the hidden layer is supplemented by a layer with the same number of units, called *context layer*, where the activation of the hidden layer on the current input are mixed with the residual values derived from the previous input. At the beginning of the sequence a constant value is assumed for the context. As long as the sequence progresses in time, the hidden layer receives both the current input, and the "context" that keeps memory of what the hidden layer has received in the past.

Recurrent networks are capable of endowing models with memory of past words in a sentence, however they are effective only for short and simple sentences. The difficulty is in maintaining memory of words that are distant but semantically and syntactically related with the current one, and forgetting words that are not necessary any more. Several improvement over Elman's recurrent neural networks have been investigated, one of the most

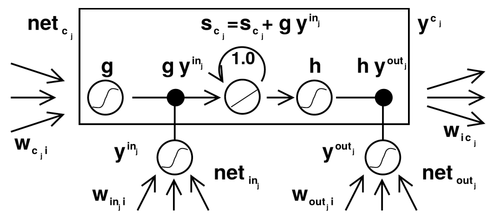


Figure 1.5: A LSTM memory neuron. In addition to its content, the neuron has several "gates" that learns to open and close access to the content, for the tasks of updating the content, or reading out the content.

popular is the Long Short-Term Memory (LSTM) proposed by Hochreiter and Schmidhuber (1997), shown in Fig. 1.5.

1.1.3 The Transformer Architecture

Despite improvements, the difficulty of neural model to capture relations between words in long sentences, and even worst across sentences, have kept DL far behind the linguistic performance of human beings. A fundamental breakthrough came with the introduction of the Transformer architecture by the Google teams (Vaswani et al., 2017).

It retains two of the solutions of the issues posed by language to neural networks, discussed previously: the word embedding coding of words into numerical vectors, and the autoencoder strategy. Its main novelty is that, unlike all previous DL models for language, it is not recurrent any more. All words are presented simultaneously as input. The overall Transformer is shown in Fig. 1.6.

The Transformer replaced recursion with a very different mechanisms, called *attention*, wich key components are shown in Fig. 1.7 The scope of attention is to encode relations between words in a sentence, binding together words that are syntactically related. The mechanism is based on a number of ancillary variables, each word in the sentence is associated with a query vector; a key vector; a value vector; and a score scalar. The query, key, and value, are computed from the word embedding, by multiplication with separated matrices. Then, for each word the score values are computed with respect to all words in the sentence, itself included. The Fig. 1.7 shows the computation steps for just two words, for simplicity. The score scalars are used to weight the values of all words, and finally the attention value for a word is the sum of all the weighted value vectors.

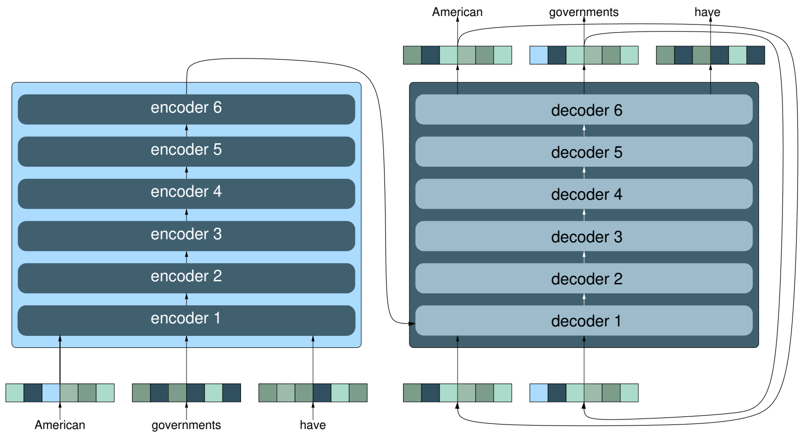


Figure 1.6: The overall Transformer architecture, with a stack of 6 encoders on the left, and a stack of 6 decoders on the right. The entire sentence is presented to the encoder, the decoder outputs one word at time, taking as input the encoder output, and the previous outputs of the decoder itself.

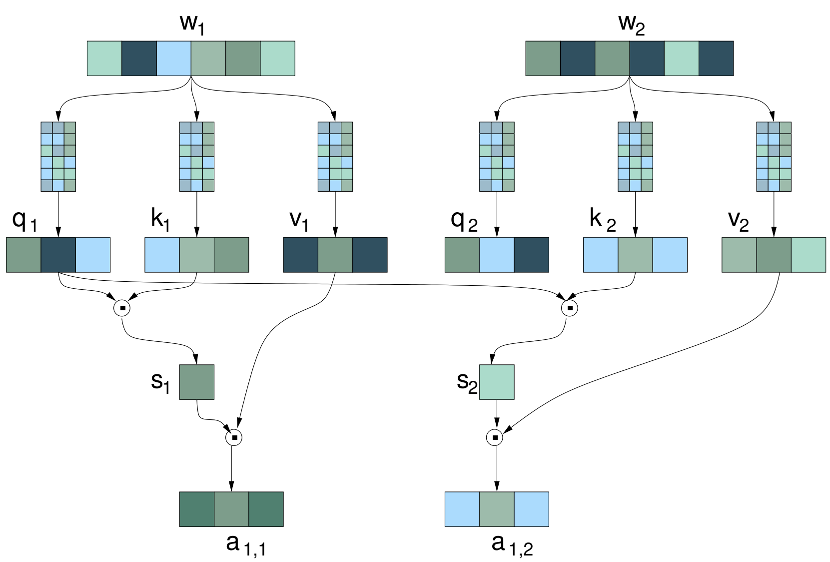


Figure 1.7: Detail of the attention mechanism of the Transformer architecture. The symbols used in the figure are the following: w = word embedding; k = key; q = query; v = value; s = score; a = attention vector. Two words only are shown, identified by the subscripts 1 and 2.

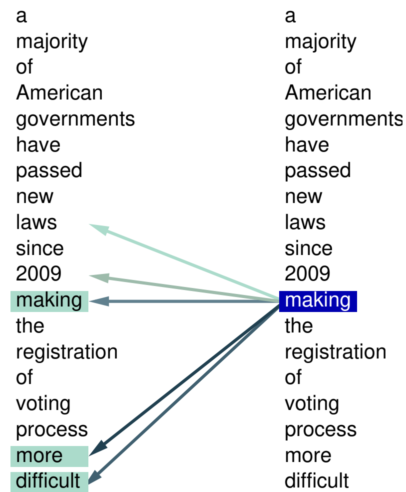


Figure 1.8: Example of the attention mechanism scores, computed for the word `making`. In addition to pointing to itself, attention correctly points with high score to `more` and `difficult`.

An example of attention scores is shown in Fig. 1.8, this is one of the original examples in (Vaswani et al., 2017), on the sentence we already quoted before. The word `making` has—correctly—the highest attention score with `more` and `difficult`, a result unlikely to be achieved with any kind of recursive network.

The names "query", "key", and "value", although evocative, are nothing more than abstractions in a clever computation of word relations in a sentence, and one should not expect the role of these variables to resemble the use of these words in everyday language. The term "attention" also needs to be taken with a grain of salt; this point will be addressed shortly after, when discussing the broader issue of how Transformer models relate with brain mechanisms.

The turning point of Transformer boosted DL for language, with a rapid progression in few years. The BERT (*Bidirectional Encoder Representations from Transformers*) model by (Devlin et al., 2019), as the name reveals, applies attention in the decoder section to both the left and the right side of the output sentence. A similar approach is followed by OpenAI's GPT (*Generative Pre-trained Transformer*) Brown et al., 2020, with the simplification of discarding the encoder part, being composed by decoders only. All current NLMs, including those here investigated for application in education, are

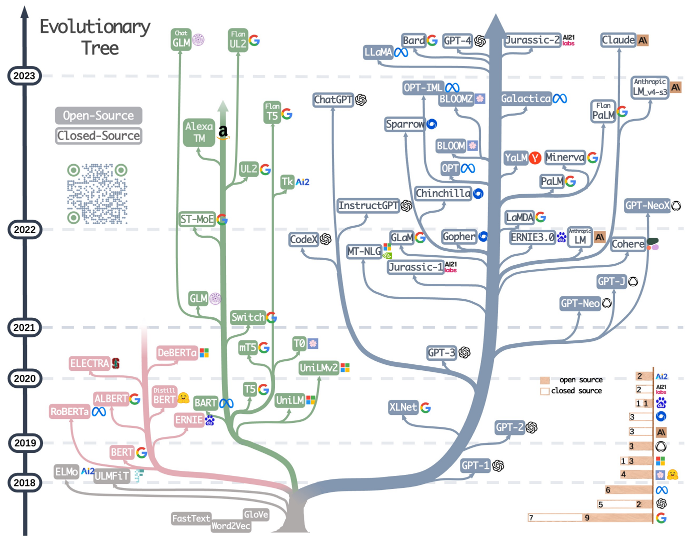


Figure 1.9: Language Models depicted as species evolved from their Transformer anchestor, by Yang et al. (2023).

Transformer based. Someone (Yang et al., 2023) has seen the Transformer like a constantly evolving species, depicting its kind of evolutionary tree, shown in Fig. 1.9.

1.2 Chatbots in Education

In this section, the use of chatbots in educational settings prior to the emergence of Language Models is explored. Limitations, advantages, and applications of chatbots in education are discussed. This section then focuses on the initial adoption and implementation of Language Models in educational environments, discussing case studies, pilot projects, or early experiments where Language Models were used in classrooms. The successes, challenges, and lessons learned from these experiences, are highlighted.

1.2.1 Chatbots before Language Models

The use of AI in education has a relatively short history but has rapidly gained popularity in recent years (Ausat et al., 2023). The first examples of AI in education can be traced back to the 1960s and 1970s when researchers began using computers to provide personalized instruction and adapt to the needs of individual students (Zhai, 2022). This could be done through chatbots, a tool that combines AI and natural language processing or other technology, enabling it to interact at a certain level of conversation with a human interlocutor through text or voice (Bishop, 2021). The idea of a chatbot can, of course, be understood as the creation of software that attempts the famous Turing Test (Turing, 1950), which in itself, however, primarily represented an abstract ideal of an intelligent machine. The endeavor to implement software capable of conversation proved to be extremely challenging, and was not even considered in the early days of AI, which had to contend with the immense difficulty of handling human language. The discipline dealing with this challenge is knkown as Natural Language Processing (NLP), and dates back to 1952 when Weaver and Bar-Hillel presented a text translation machine at MIT. Subsequently, throughout the 1960s, all efforts were directed in this direction, leading to the founding of the AMTCL (Association for Machine Translation and Computational Linguistics) in Princeton in 1962. Projects for automatic translation machines continued with GAT-SLC (Georgetown Automatic Translation-Simulated Linguistic Computer) by Zarenchnak and Brown, 1961 for translation from Russian to English. These early researches did not yield particularly notable results, so much so that in November 1966, the ALPAC (Automatic Language Processing Advisory Committee) advised the United States government to stop supporting Machine Translation. It was evident that before moving on to automatic translation, it was necessary to work on the fundamentals of natural language (Earley, 1970; Kasami, 1965). In between, there was a sort of scandal, the well-known ELIZA program by Weizenbaum, 1966, which conversed with a user by interpreting the role of a psychotherapist. It had, at the same time, enormous public success and was disdained by the NLP community. Despite its very limited conversational abilities, ELIZA was able to effectively simulate the role of a Rogerian psychologist, who encourages the patient to reflect on their own responses without delving too deeply into details and ignoring all references to the real world. Despite its success, ELIZA was based on a simple logic of triggering predefined responses activated by keywords (Brandtzaeg & Følstad, 2017). Even with all the discredit by the Community NLP, ELIZA served as inspiration for the subsequent development of other chatbots (Klopfenstein et al., 2017; Pérez et al., 2020).

Another step forward in the history of chatbots was the creation of ALICE (Artificial Linguistic Internet Computer Entity) in 1995, the first online chatbot inspired by ELIZA (Wallace, 2008). ALICE relied on pattern matching without actual perception of the entire conversation, but it had the ability to engage in web-based discussions that allowed for longer conversations and covered a wide range of topics (das Graças Bruno Marietto et al., 2013). However, it took several years for it to be improved to win the Loebner Prize for the most human-like computer program (Bradeško et al., 2017). ALICE was developed with a new language created for this purpose, Artificial Intelligence Markup Language (AIML), which is the most critical difference between ALICE and ELIZA. ALICE's Knowledge Base consisted of approximately 41,000 patterns and related templates, a significant increase compared to ELIZA, which only had 200 keywords and rules. However, ALICE did not have intelligent features and could not generate human-like responses expressing emotions or attitudes. During the 1960s-1970s, Eliza and ALICE were not the only forms of AI invented for human-AI interaction. PLATO (Programmed Logic for Automatic Teaching Operations) was developed at the Coordinated Science Laboratory of the University of Illinois to explore the possibilities of automation in individualized instruction (Lyman, 1972). PLATO aimed to provide interactive and personalized instruction to students through the use of computer terminals. The system involved a central mainframe computer connected to multiple terminals distributed in classrooms and other locations. Each terminal consisted of a screen, a keyboard, and often a light pen for input. Another significant development was "Coach: The Virtual Tutor," created by Carnegie Mellon University. Coach was designed to assist students with math assignments and could respond to student questions by providing explanations and assistance in solving math problems. It was a significant step in the application of chatbots for educational purposes and paved the way for further developments and implementations in the field. These works have laid the foundations for the development of more advanced educational tools and AI-based platforms that have become common in today's classrooms. In the 1980s and 1990s, the proliferation of personal computers and the internet enabled the development of more sophisticated educational software and online learning platforms. These early efforts largely focused on providing content and assessments rather than delivering personalized instruction or adapting to the individual needs of students (Zhai, 2022). In the subsequent years, between 1970 and 2000, we can observe a tremendous progression of chatbots. It is important to note that each work for a specific project evolved and built a fundamental space for future generations. In short, some of the important chatbots developed prior to the era of language models used for educational purposes include:

- TALE-SPIN: a narrative generation chatbot for educational purposes (1970s). Developed by James Meehan in the 1970s, it was a natural language understanding system that used story comprehension as the basis for conversation. Although not exclusively designed for educational purposes, TALE-SPIN's conversational capabilities were used in educational applications.

- SCHOLAR: developed in the late 1980s by researchers at Yale University. Developed in the late 1980s at the University of Southern California, it was an early intelligent tutoring system. It incorporated a chatbot component that allowed students to interact with the system using natural language to receive personalized feedback and assistance.

- AutoTutor: emerged in the 1990s, created by researchers at the University of Memphis. Developed in the 1990s at the University of Memphis, it was an intelligent tutoring system designed to simulate human tutoring interactions. It employed a conversational agent that provided interactive tutoring in various subjects.

- ASK Systems: developed in the late 1990s by Professor Tom Murray at the University of Massachusetts Lowell. Developed in the late 1990s, it was a series of intelligent tutoring systems designed to teach computer programming languages. The ASK chatbot component allowed students to ask questions and receive personalized feedback and guidance.

- ALEKS (Assessment and Learning in Knowledge Spaces): created in the late 1990s by a team of mathematicians, cognitive scientists, and software engineers. Developed in the late 1990s, it is an adaptive learning system that incorporates a chatbot-like interface. It provides individualized instruction and assessment in various subjects, primarily mathematics.

The experiences described so far, although important in introducing AI tools to the educational world, belong to a different era. Neural language models have marked a turning point, and both the range of potential applications and the sophistication of these tools have opened up an entirely new landscape to be explored. These new tools are able to adapt to the needs of individual students, provide personalized instruction, and even grade assignments. Today, thanks to the proliferation of NLMs, AI is being utilized in various ways in education (Rudolph et al., 2023a). Below, we will discuss in detail the risks and opportunities of new Language Model-based chatbots in education.

1.2.2 Early experiences of Language Models in the Classroom

The history of NLM can be traced back to the early explorations of the use of ANN techniques in the field of NLP, but it is in the period between 2000 and 2018 that the progress achieved begun to have impact in science and society in general. But the use of NLM for educational purposes began to gain traction in recent years only, particularly with the emergence of models like OpenAI's GPT series. The underlying technology is key to further innovations and, despite critical views and even bans within communities and regions, NLMs are here to stay (Kasneci et al., 2023). When it comes to AI in education, we refer to a process called digital transformation through which chatbots are introduced into the field of education (Pérez et al., 2020). The proliferation of chatbots based on Language Models like ChatGPT has initially raised concerns, and many authors have reported them in early scientific articles (Firat, 2023; Kasneci et al., 2023; Mollick & Mollick, 2022; Shoufan, 2023; Tlili et al., 2023). One challenge in using chatbots based on NLMs for assessment in education is the possibility of plagiarism. AI-powered automated writing systems are designed to generate essays based on given parameters or prompts. This means that students could potentially use these systems to cheat on their assignments by submitting essays that are not their own work (Dehouche, 2021). This undermines the purpose of education, which is to assess and educate students, and could devalue academic degrees.

Another challenge is the potential for chatbots to unfairly advantage some students over others. For instance, if a student has access to GPT-3 and uses it to generate high-quality written assignments, they may gain an unfair advantage over other students who do not have access to the model. This could lead to disparities in the evaluation process (Cotton et al., 2023). Given the challenges associated with evaluating student assignments written using chat APIs, there are several strategies that academic staff can employ to address these challenges (Tlili et al., 2023). Firstly, clear and detailed instructions can be provided to students on how to structure their assignments. This can help ensure that assignments are written in a more organized and coherent manner. Secondly, academic staff can use rubrics to assess the quality of students' work. This can help ensure that students' efforts and understanding of the material are accurately evaluated. Lastly, a combination of automated and manual assessment techniques can be used to evaluate students' comprehension of the material. This can help ensure that students' true level of understanding is accurately assessed.Another challenge associated with the use of chatbots based on Language Models is the presence of biases. Since these NLMs are trained on large amounts of textual data collected from the web, they can inadvertently absorb biases and prejudices present in that data. This means that chatbots could produce responses or suggestions that reflect social, cultural, or other biases present in the training text, without students being aware of it. Moreover, the use of chatbots could lead to a certain dependence on technology, with students relying solely on the answers provided by chatbots instead of developing their own critical thinking and problem-solving skills. This could limit their development of essential competencies in the learning process. To mitigate this risk, it is crucial to adopt approaches for quality assessment and control. Developers and educators need to closely monitor chatbot responses to identify and correct any biases or biases that may emerge. It is important to fine-tune and train NLM to promote equity and inclusion, avoiding the perpetuation of existing biases. Additionally, it is essential to educate students about the importance of developing critical thinking towards chatbot responses. Students should be aware that the answers provided by chatbots may not always be completely accurate or unbiased, and they should be encouraged to critically evaluate the information received and develop their own opinions and perspectives.

1.2.3 Opportunities for Learning and Teaching

After an initial phase of great concern, literature began to include articles that sought to focus on the most functional ways to use NLM in education. This is important because, despite the system's efforts to slow down the spread of this technology, the progress already made is too advanced for a pause to have a real impact (Qadir, 2022). Language Models are here to stay, and it is important to understand how to use them effectively in education. Below are summarized a series of opportunities for learning from both the students' and teachers' perspectives (Zhai, 2022). Through the utilization of NLMs, opportunities for the enhancement of learning and teaching experiences may be possible for individuals at all levels of education, including primary, secondary, tertiary, and professional development.

- For elementary school students: NLMs can assist in the development of reading and writing skills (e.g., by suggesting syntactic and grammatical corrections), as well as in the development of writing style and critical thinking skills. These models can be used to generate questions and prompts that encourage students to think critically about what they are reading and writing and to analyse and interpret the information presented to them. Additionally, NLMs can also assist in the development of reading comprehension skills by providing students with summaries and explanations of complex texts, which can make reading and understanding the material easier.

- For middle and high school students: NLMs can assist in the learning of a language and of writing styles for various subjects and topics, e.g., mathematics, physics, language and literature, and other subjects. These models can be used to generate practice problems and quizzes, which can help students to better understand, contextualize and retain the material they are learning. Additionally, NLMs can also assist in the development of problem-solving skills by providing students with explanations, step-by-step solutions, and interesting related questions to problems, which can help them to understand the reasoning behind the solutions and develop analytical and out-of-the-box thinking.

- For university students: NLMs can assist in the research and writing tasks, as well as in the development of critical thinking and problem-solving skills. These models can be used to generate summaries and outlines of texts, which can help students quickly understand the main points of a text and organize their thoughts for writing. Additionally, NLMs can also assist in the development of research skills by providing students with information and resources on a particular topic and hinting at unexplored aspects and current research topics, which can help them to better understand and analyse the material.

- For group and remote learning: NLMs can be used to facilitate group discussions and debates b providing a discussion structure, real-time feedback, and personalized guidance to students during the discussion. This can help to improve student engagement and participation. In collaborative writing activities, where multiple students work together to write a document or a project, language models can assist by providing style and editing suggestions as well as other integrative co-writing features. For research purposes, such models can be used to span the range of open research questions in relation to already researched topics and to automatically assign the questions and topics to the involved team members. For remote tutoring purposes, they can be used to automatically generate questions and provide practice problems, explanations, and assessments that are tailored to the student's level of knowledge so that they can learn at their own pace.

- To empower learners with disabilities: NLMs can be used in combination with speech-to-text or text to- speech solutions to help people with visual impairment. In combination with the previously mentioned group and remote tutoring opportunities, language models can be used to develop inclusive learning strategies with adequate support in tasks such as adaptive writing, translating, and highlighting important content in various formats. However, it is important to note that the use of NLMs should be accompanied by the help of professionals such as speech therapists, educators, and other specialists that can adapt the technology to the specific needs of the learner's disabilities.

- For professional training: NLMs can assist in the development of language skills that are specific to a particular field of work. They can also assist in the development of skills such as programming, report writing, project management, decision-making, and problem-solving. For example, NLMs can be fine-tuned on a domain-specific corpus (e.g. legal, medical, IT) in order to generate domain-specific language and assist learners in writing technical reports, legal documents, medical records etc. They can also generate questions and prompts that encourage learners to think critically about their work and to analyse and interpret the information presented to them.

In conclusion, Language Models have the potential to provide a wide range of benefits and opportunities for students and professionals at all stages of education. They can assist in the development of reading, writing, math, science, and language skills, as well as provide students with personalized practice materials, summaries, and explanations, which can help to improve student performance and contribute to enhanced learning experiences. Additionally, NLM can also assist in research, writing, and problem-solving tasks, and provide domain-specific language skills and other skills for professional training. However, as previously mentioned, the use of these models should be done with caution, as they also have limitations such as lack of interpretability and potential for bias, and unexpected brittleness in relatively simple tasks which need to be addressed.

Chatbots based on current NLM, such as ChatGPT, have the potential to revolutionize teaching and assist in teaching processes. Below we provide only a few examples of how these models can benefit teachers:

- For personalized learning: teachers can use NLM to create personalized learning experiences for their students. These models can analyse students' writing and responses and provide tailored feedback and suggest materials that align with the student's specific learning needs. Such support can save teachers time and effort in creating personalized materials and feedback and allow them to focus on other aspects of teaching, such as creating engaging and interactive lessons.

- For lesson planning: NLM can also assist teachers in the creation of (inclusive) lesson plans and activities. Teachers can input into the models the corpus of documents based on which they want to build a course. The output can be a course syllabus with a short description of each topic. Language models can also generate questions and prompts that encourage the participation of people at different knowledge and ability levels and elicit critical thinking and problem-solving. Moreover, they can be used to generate targeted and personalized practice problems and quizzes, which can help to ensure that students are mastering the material.

- For language learning teachers of language classes can use NLM in an assistive way, e.g., to highlight important phrases, generate summaries and translations, provide explanations of grammar and vocabulary, suggest grammatical or style improvements, and assist in conversation practice. Language models can also provide teachers with adaptive and personalized means to assist students in their language-learning journey, which can make language learning more engaging and effective for students.

- For research and writing: NLM can assist teachers of university and high school classes to complete research and writing tasks (e.g., in seminar works, paper writing, and feedback to students) more efficiently and effectively. The most basic help can happen at a syntactic level, i.e., identifying, and correcting typos. At a semantic level, NLM can be used to highlight (potential) grammatical inconsistencies and suggest adequate and personalized improvement strategies. Going further, these models can be used to identify possibilities for topic-specific style improvement. They can also be used to generate summaries and outlines of challenging texts, which can help teachers and researchers to highlight the main points of a text in a way that is helpful for further deep dive and understanding of the content in question.

- For professional development: NLM can also assist teachers by providing them with resources, summaries, and explanations of new teaching methodologies, technologies, and materials. This can help teachers stay up to date with the latest developments and techniques in education and contribute to the effectiveness of their teaching. They can be used to improve the clarity of the teaching materials, locate information or resources that professionals may need as they learn on the job, as well as used for on-the-job training modules that require presentation and communication skills.

- For assessment and evaluation: teachers can use NLM to semi-automate the grading of student work by highlighting potential strengths and weaknesses of the work in question, e.g., essays, research papers, and other writing assignments. This can save teachers a significant amount of time for tasks related to individualized feedback to students. Furthermore, Language Models can also be used to check for plagiarism, which can help to prevent cheating. Hence, a NLM can help teachers to identify areas where students are struggling, which adds to a more accurate assessment of student learning development and challenges. Targeted instruction provided by the models can be used to help students excel and to provide opportunities for further development.

- The acquaintance of students with AI challenges related to the potential bias in the output, the need for continuous human oversight, and the potential for misuse of NLMs are not unique to education. In fact, these challenges are inherent to transformative digital technologies. Thus, we believe that, if handled sensibly by the teacher, these challenges can be insightful in learning and education scenarios to acquaint students early on with potential societal biases, and risks of AI application.

In conclusion, NLMs have the potential to revolutionize teaching from a teacher's perspective by providing teachers with a wide range of tools and resources that can assist with lesson planning, personalized content creation, differentiation and personalized instruction, assessment, and professional development. Overall, NLMs have the potential to be a powerful tool in education, and there are several ongoing research efforts exploring its potential applications in this area. (Kasneci et al., 2023). Chatbot applications can provide students with immediate feedback and tailored responses to their questions. AI can also be used to personalize the learning experience by providing recommendations for resources, such as books and websites, that are tailored to the student's needs and interests. Chatbot applications can also provide educational resources, such as study guides and lecture notes, to help students better understand the material (Dwivedi et al., 2023). Based on practical experiences use of chatbots for educational goals is beneficial.

Chapter 2

Critical Thinking and ways to enhance it

This chapter introduces critical thinking (CT), a psychological construct that represents an important 21st- century skill and a significant learning objective in education (Y. Liu & Pásztor, 2022). In an era increasingly shaped by the rapid circulation of information—often ambiguous, biased, or misleading—CT offers individuals the cognitive tools to evaluate information critically, form evidence-based beliefs, and make reasoned decisions (Ollinheimo & Hakkarainen, 2023a; Papp et al., 2014). This is particularly relevant in the context of digital environments, where information quality is highly variable and persuasive content can easily distort judgments (Fabio & Suriano, 2023). Beyond its societal relevance, CT is also foundational in science and technology, where it supports the rigorous analysis of data and helps filter out misinformation (Huang et al., 2015; Sahoo & Mohammed, 2018). In education, CT encompasses both analytical skills—such as reasoning and argumentation—and reflective capacities, such as questioning assumptions and evaluating sources within real-life and ethical contexts. Despite the recognized importance of CT, assessing it remains a challenge. Many tools currently used suffer from narrow scopes or inadequate psychometric properties.Moreover, the teaching and enhancement of CT require solid theoretical and methodological approaches, capable of adapting to diverse educational, cultural, and technological contexts.

This chapter is divided into two main sections. The first section outlines the landscape of CT in the human domain, beginning with a conceptual definition of the construct. It presents the most relevant theoretical models, followed by an overview of assessment tools, with particular attention to critical reasoning tests. Finally, it explores strategies and approaches for teaching CT, emphasizing its importance in contemporary educational settings. The second section introduces an emerging field: the relationship between CT and AI, with particular reference to the so-called machine psychology that will be addressed in depth in §2.2.1. This part discusses whether and how advanced systems such as ChatGPT can be considered critical thinkers, and reflects on the potential role of AI in promoting, simulating, or even reshaping human CT processes. The overall aim of the chapter is thus twofold: on one hand, to provide an updated and structured overview of human CT; on the other, to open a space for reflection on its possible extensions within the artificial domain.

2.1 Mapping the Landscape of Critical Thinking

2.1.1 What is critical thinking?

CT represents a complex and essential set of cognitive skills that play a crucial role in the ability to form informed judgments (C. M. Evans, 2020; Hwang et al., 2023). These skills are foundational for success in learning and professional careers and are particularly vital in a world that is constantly changing and evolving (Alharbi et al., 2022; Y. Liu & Pásztor, 2022; Thornhill-Miller et al., 2023). In an era where technology plays a central role in all fields, CT provides the necessary tools to rigorously and rationally evaluate the vast amount of information to which we are exposed, enabling us to pursue DL and make informed decisions (Fabio & Suriano, 2021; Meirbekov et al., 2022). The ability to think critically is crucial for addressing the increasingly intricate and complex challenges that characterize contemporary society (Santos-Meneses et al., 2023).

Different authors have provided definitions of the CT construct. According to Ennis (2002), CT represents the ability to impartially analyze a problem, examine various perspectives, and arrive at conclusions based on solid evidence. Halpern (2006) emphasizes the importance of being open to new ideas and the willingness to explore them in depth, avoiding seemingly obvious solutions to achieve deeper results. Elder and Paul (2010) define CT as an intellectual process involving the analysis, evaluation, and improvement of one's own thinking, with the goal of making more informed decisions and solving problems more effectively. This process is based on careful exploration, evaluation, and reflection on ideas and arguments (Moore, 2010). Brookfield and Preskill (2016) describe CT as the ability to critically examine one's beliefs and assumptions, along with the ability to consider different perspectives. O. L. Liu et al. (2014) have identified common elements in many of these definitions, including skills such as evidence evaluation, argument analysis, and drawing conclusions based on evidence. The complex nature of this construct contributes to the difficulty of providing a singular and exhaustive definition of CT. The lack of a clear operational definition represents one of the main challenges in assessing this crucial cognitive skill.

2.1.2 Evaluating critical thinking performance

Many studies on CT have relied on self-assessment methods (R. J. Shavelson et al., 2019), such as the Critical Thinking Disposition Scale (CTDS) developed by Sosu (2013) and the Marmara Critical Thinking Dispositions Scale developed by Özgenel and Çetin (2018). Despite their widespread use, such tools focus on participants' perceptions of their own CT abilities or their inclination to use them, rather than directly measuring their actual skills (Braun et al., 2020). Individual perceptions can vary significantly from actual performance; subjects may exaggerate or underestimate their abilities (Bowman, 2010). Consequently, it is essential to adopt performance-based assessments involving complex tasks that evoke real-life situations, allowing for a more accurate evaluation of CT abilities (Davey et al., 2015; Hyytinen & Toom, 2019; R. Shavelson, 2010). Performance-based assessments can take various forms, ranging from open-ended tasks to multiple-choice tasks (Fajari et al., 2021; Kember et al., 2000; Messick, 1994; Zuriguel-Pérez et al., 2017). Among these, one of the most widely used is the California Critical Thinking Skills Test (CCTST; Lin et al., 2018), which consists of 52 multiple-choice items that assess analysis, evaluation, deduction, inductive reasoning, and inferential reasoning. Another extensively used tool for CT performance assessment is the Halpern Critical Thinking Assessment (HCTA) developed by Halpern (2013), which evaluates five categories of CT, including verbal reasoning, argument analysis, and the ability to make informed decisions through the presentation of 25 everyday scenarios. While these tools are widely used for their convenience, they have the drawback that it is possible to guess or eliminate incorrect answers without a deep understanding of the subject matter (Hyytinen et al., 2015, 2020). Consequently, responses to these tasks provide limited information about underlying thinking processes and often focus on individual skills, overlooking the complex nature of CT, which open-ended performance-based tasks tend to capture (Hyytinen et al., 2021). Although open-ended performance-based tasks may be seen as expensive and potentially subject to subjective interpretations in scoring, this challenge can be addressed through rigorous rater training and the implementation of detailed assessment criteria (Borowiec & Castle, 2019). Finally, many established tools produce scores that are difficult to interpret because they are not explicitly rooted in a CT theory, making the score's meaning vague or ambiguous (O. L. Liu et al., 2014). Therefore, there is a clear need for a valid, easy-to-assess tool that is grounded in theory.

2.1.3 Teories applicable to critical thinking

There are different models that seek to explain CT. The Reflective Judgment Model (King & Kitchener, 1985) describes the development of CT ability through three consecutive stages. The pre-reflective stage is characterized by a tendency to rely on direct sources of information rather than critically evaluating evidence, and the belief that all problems are clearly defined and solvable. In the quasi-reflective stage, there is a degree of uncertainty and partial evaluation of evidence to reach a non-ambiguous conclusion. Reflective reasoning involves both accepting uncertainty and a clear decision-making process; while maintaining an open conclusion to further evidence, reflective thinkers express judgments based on evidence (King & Kitchener, 1994). Another model suitable for explaining CT is the dual-process model (Stanovich & West, 2000b), which highlights how reasoning occurs through two distinct modes, providing a more comprehensive framework for the cognitive dynamics involved in critical decisions and judgments in complex situations. The dual-process model, applied to critical thinking theory, emphasizes the presence of two distinct reasoning systems: System 1 and System 2 (J. S. Evans & Stanovich, 2013). System 1 operates through automatic and intuitive mental processes guided by information processing heuristics (Frankish, 2010; Osman, 2004). In these situations, decisions and judgments are formulated instantly without the conscious involvement of the individual, relying on prior knowledge and beliefs (Fabio & Suriano, 2025; Okuhara et al., 2020). In contexts that favor the use of System 1, with its propensity for heuristic use, erroneous assessments may emerge. An example could be: "Drug abuse is a choice. It is not a disease; drug users choose it." The person responding is using their pre-existing belief that is part of the automatic systems. On the other hand,System 2 operates through analytical information processing, requiring deliberate control by the individual (Sowden et al., 2019a). Although this system is slower and demands more cognitive resources, it enables a more logical processing of information, in line with the standards of critical thinking theory (C. Zhao & Hu, 2021). This system, positively associated with CT, facilitates the formulation of more accurate judgments and opinions, source evaluation, and openness to considering multiple perspectives (Fabio & Suriano, 2023). An example could be: "It seems that many factors are involved. I think there might be genetic factors at play, as seen in some families where different members have access to drugs. Part of it could be linked to the first time an individual decides to take drugs. Unfortunately, I believe that repeated exposure to substances leads to brain changes, limiting the individual's freedom of choice due to addiction."

The integration of King and Kitchener's (1985) Reflective Judgment Model and Evans and Stanovich's (2013) Dual Process Model provides a comprehensive conceptual framework for understanding the dynamic process of CT, especially in complex decision-making contexts. Initially, during the pre-reflective phases, individuals inclined toward automatic judgments may exhibit a preference for System 1, characterized by intuitive mental processes and heuristic-based reasoning. This may result in quick and superficial decisions, often influenced by prior knowledge and intuitive perceptions. However, progress through the stages of reflective judgment may indicate a transition toward increased awareness and activation of System 2. In this phase, individuals can apply analytical information processing, exercising deliberate control, initially partial and subsequently complete, to avoid cognitive biases and achieve higher levels of CT. In the examples mentioned above, the same examples are used to interpret them with both the dual-process and reflective models. In summary, the integration of these models suggests that the developmental path of CT involves a gradual transition from the predominant use of System 1 in pre-reflective stages to a more conscious and balanced adoption of System 2 in the later stages of reflective judgment. This transition involves a phase where controlled processes are not fully applied but only partially. This evolution enables individuals to overcome the limitations of automatic judgments, mitigating cognitive biases and facilitating the attainment of higher levels of CT in complex contexts.

2.1.4 Critical Reasoning Assessment

An open-response tool for assessing CT skills is the Critical Reasoning Assessment (CRA) developed by Anghel et al. (2021b), based on the Reflective Judgment Model (King & Kitchener, 1985). Unlike most tools, CRA measures CT skills in more ecological contexts, such as moral or ethical dilemmas, rather than narrower academic or formal settings. The tool is designed to assess critical thinking abilities based on participants' performance rather than self-reported information. Note that for these assessment tools, it is more appropriate to refer to "critical reasoning" rather than "critical thinking." As discussed above, critical thinking encompasses a range of human mental dispositions. Evaluating these necessarily goes beyond simply measuring the possession of such dispositions or the circumstances in which they are manifested, since we only have access to the subjects' written output, where only the performance aspect of CT is exercised. The structure of CRA is built upon the Reflective Judgment Interview (King & Kitchener, 1994), utilizing specially designed dilemmas to stimulate CT and in-depth reflection on key themes such as ethical decision-making and complex problem-solving. The choice of CRA for this study is grounded in its unique focus on real-world scenarios, aligning with the broader and more practical applications of critical thinking skills. Moreover, CRA's emphasis on performance-based assessment addresses concerns about the limitations of multiple-choice tasks, providing a more comprehensive understanding of participants' underlying thinking processes. By utilizing open-ended tasks, CRA captures the complexity of CT, offering a nuanced evaluation that goes beyond the constraints of other widely used tools. This aligns with the necessity to employ an assessment tool that not only measures critical thinking skills but also provides meaningful insights into the depth and breadth of individuals' cognitive abilities. The structure of the CRA utilizes specially designed dilemmas to stimulate CT and in-depth reflection on key themes such as genetics vs. choice,fairness, and compassion. An example related to genetics vs. choice is:

> Some researchers argue that drug abuse is, at least in part, due to genetic factors. They refer to the results of numerous studies that support this claim. Other researchers, however, believe that drug abuse is a choice, asserting that the reason why different family members often engage in drug abuse is due to sharing family experiences, socio-economic status, or common employment.

Participants undergoing the CRA are invited to respond to a series of targeted questions related to each dilemma, aiming to explore their decision-making process and investigate the foundations of their perspective. This approach highlights not only the adopted position but also the logical basis and arguments that could support alternative viewpoints. The questions aim to capture five dimensions: cognitive complexity, reasoning style, openness, nature of knowledge, and nature of justification.

Cognitive complexity involves the ability to assess the pros and cons of an issue or choice and understand that multiple levels of analysis of a dilemma are possible.

Question: What is your opinion about this issue?

Low score example: "Drug abuse is a choice. It is not a disease; drug users choose it."

High score example: "It seems that many factors are involved. I think there might be genetic factors at play, as seen in some families where different members have access to drugs. Part of it could be linked to the first time an individual decides to take drugs. Unfortunately, I believe that repeated exposure to substances leads to brain changes, limiting the individual's freedom of choice due to addiction."

Reasoning Style assesses the ability to use evidence systematically to start from a hypothesis and arrive at a conclusion, thereby manifesting a justifiable solution or worldview. *Question:* How have you come to hold this point of view? On what do you base that point of view? An example of a low score would be if the individual responds: "It is my opinion based on observing the people in my social network that they abuse drugs." On the other hand, a high score is attributed when the individual states: "I have reasoned through this; the knowledge I have acquired at university leads me to conclude that both factors contribute. Moreover, research based on sociological and medical results seems to indicate both factors."

Openness refers to the awareness and understanding of others' perspectives, as well as the recognition of the limitations of one's own worldview, without presupposing that one's perspective is universal or absolute. *Question:* Can you ever know for sure that your position on this issue is correct? Why or why not? When two people differ about matters such as this, is it the case that one opinion is right, and one is wrong? If yes, what do you mean by "right"? If no, can you say that one opinion is in some way better than the other? What do you mean by better? An example of a low score would be if the individual responds: "Everyone forms their own opinions, and I continue to think it's a choice." On the other hand, a high score is attributed when the individual says: "I avoid using the term 'certainty' because I believe it's important to maintain a margin of doubt, considering our often-partial knowledge. I assess opinions based on their alignment with empirical data and consider potential cognitive biases. In some situations, it might be appropriate to speak in terms of right/wrong, but I don't consider this dichotomy exhaustive. For these reasons, I believe that both genetic and environmental factors come into play."

Nature of Knowledge involves how an individual assesses and applies knowledge, considering sources, reliability, and relevance of information in decision-making. *Question:* How is it possible that people have such different points of view about this subject? An example of a low score is given when the subject responds: "Because often there is a tendency not to fully understand the reality in which we live today, unfortunately." On the other hand, a high score is assigned when the subject says: "I base what I told you on my personal experiences, on the assessment that there can be errors or heuristics when we choose, personality factors, different modes of thinking, and varying analytical tendencies."

Nature of Justification examines how an individual uses concepts, evidence, and expert opinions to justify a worldview or point of view, emphasizing the process of providing a rational basis for one's beliefs. *Question:* How is it possible that experts in the field disagree about this subject? A low score example is when the subject responds: "I think I don't need to justify my beliefs." In contrast, a high score is given when the subject says: "Complex phenomena often involve intricate cause-and-effect relationships and debates on co-factors. Different research methods lead to various conclusions, making unanimous consensus rare. The same studies on the drugs we are talking about have demonstrated this. The results can generate conflicting data, complicating the attribution of importance to a specific cause."

The scoring is analytical in nature, meaning that each of the five dimensions is evaluated separately (complexity, reasoning, openness, nature of knowledge, and nature of justification). The original scoring scale by Anghel et al. (2021b) consisted of 13 points (7 main levels plus intermediate scores), whereas the standardized version by Fabio, Plebe, Ascone, and Suriano (2025) employs a 7-point Likert scale ranging from 1 to 7. In the standardized version, the rubric was revised to improve clarity, consistency, and practical applicability of the assessment, thereby ensuring greater reliability and ease of use. As a result, this test provides an extremely detailed scoring profile with precise indications for each phase. The scoring schemes for both versions are presented in Tables 2.1 and 2.2, respectively.

Table 2.1: The original CRA scoring rubric

	Cognitive Complexity	Reasoning Style	Openness	Nature of Knowledge	Nature of Justification
1	Knowledge is concrete, tangible, singular, & observed; clichéd.	Opinion = fact; no reasoning	Uninformed & naive; deliberately ignores points of view that are unlike their own	Knowledge needs no justification; right or wrong is known; no other options	Alternatives do not exist; evidence is not evaluated; knowing is egocentric
1.5	Knowledge is concrete; alt. views emerging; not qualified as right or wrong	Reasoning is attempted, but facts are inaccurate	Still naive & uninformed, recognizes but dismisses other points of view	Knowledge is certain, right or wrong; authorities justify knowledge	Justification combines egocentrism and with what subject has been told by authorities

Table 2.1: The original CRA scoring rubric

	Cognitive Complexity	Reasoning Style	Openness	Nature of Knowledge	Nature of Justification
2	Knowledge is right or wrong; solutions are simple & easy	Reasoning is illogical; opinion & evidence are blurred	Other points of view are possible but wrong	Knowledge is absolutely certain or un-certain; different views are just wrong	Justifies beliefs via authorities; not based on evidence but what subject has been told
2.5	Knowledge is right or wrong but uncertainty prevents choosing right or wrong	Reasoning is illogical; starting to see difference between opinion & evidence	Open to other points of view but they exist in limbo without right or wrong designation	Knowledge may be uncertain; moving away from labeling views as right or wrong.	Acknowledges good & bad authorities; uses personal experience in combination with good authority to justify decision
3	Knowledge is true, false or uncertain; ambiguity is troubling	Some logic, that is personal & subjective; what feels right	Some openness; rejects beliefs rather than be uncertain	Knowledge will eventually be known; one answer is just as good as another	Decisions are tentative; fact & opinion are different; authority is questioned
3.5	Knowledge is questioned; emerging toleration for ambiguity	Reasoning appears more logical but does not pose any point of view	Open to other points of view beginning to accept uncertainty	Knowledge may not be eventually known; beginning to consider that knowledge is uncertain	Moving toward having a strong opinion but still subjective based on unexamined facts & opinions
4	Knowledge is uncertain; issues are complex but without sub-issues	Beginning to realize role of evidence; but does not consistently argue with evidence	Open to other points of view; stubborn or wishy-washy; contradictory	Knowledge is idiosyncratic, abstract, uncertain, relativistic; discrete differences; no gray	Expresses strong point of view without objectivity; evidence is incomplete; authority is dogmatic
4.5	Knowledge is complex; certainty & realization of sub-issues are emerging but limited	Reasoning is logical; argument is somewhat explicit with limited evaluation of evidence	Open to other points of view; objective about some/subjective about others.	Knowledge is idiosyncratic and relativistic but not black or white; some ambiguity	Point of view is expressed with some objectivity, but evidence is incomplete
5	Knowledge is complex; experience is limiting; evidence has many sides	Reasoning is logical; with explicit & consistent evaluation of evidence	Sees diverse points of view; objective about all points of view	Knowledge is domain-specific, uncertain, & complex; difference relates to different worldview	Justifies point of view based on evidence that is judged qualitatively using simple/learned rules of inquiry; offers balanced "big picture"
5.5	Knowledge is complex; emerging exploration & analysis	Reasoning is logical; recognizes that some evidence is more compelling; beginning to articulate such claims	Examines many points of view; dismisses some unreasonable claims	Knowledge is complex & contextual; limited comparison of points of view across domains	Justifies point of view based on evidence but starting to inquire & compare perspectives
6	Knowledge is complex & analyzed across points of view	Reasoning is logical based on evaluation of evidence; articulates more compelling claims	Examines many points of view; dismisses unreasonable claims, but offers no final personal view	Knowledge is judged qualitatively & although valid, not totally defensible; considers credibility of claims	Assumes but does not construct point of view; evaluates strength of evidence & experts' claims
6.5	Knowledge is analyzed & synthesized, but not con-structed to yield a perspective	Reasoning is logical; emerging strategy; with some abstraction	Dismisses unreasonable claims; close to offering personal view	Knowledge is judged qualitatively; some perspectives could be more defensible than other; more inquiry is needed	Beginning to construct point of view using strength of evidence and claims; beginning to synthesize across domains

Table 2.1: The original CRA scoring rubric

	Cognitive Complexity	Reasoning Style	Openness	Nature of Knowledge	Nature of Justification
7	Knowledge is complex, analyzed, synthesized, constructed to form a coherent perspective	Reasoning is logical, strategized, generalized into abstractions supported by evidence	Sees why others hold points of view but owns personal view; is open to new information	Knowledge results from rigorous inquiry across multiple perspectives, across multiple contexts & is more or less reasonable & defensible	States opinions firmly based on evaluated evidence; abstracts across & within domains; constructs higher-order thinking

The Critical Reasoning Assessment (CRA) was adapted and validated for the Italian population through a multi-phase process. A pilot test involving 79 participants addressed initial issues and confirmed the CRA's ability to differentiate between high and low performers. The first study, conducted with 123 participants, validated the CRA's unidimensional structure and its consistency in measuring critical reasoning. A subsequent study with 480 participants further supported this structure, demonstrating strong fit indices.

The adaptation process included translation and back-translation by bilingual experts to ensure linguistic accuracy. The CRA showed excellent reliability (Cronbach's $\alpha = 0.93$) and strong convergent validity, evidenced by positive correlations with the Critical Reasoning Disposition Inventory and academic performance, as well as divergent validity, demonstrated by negative correlations with dysfunctional beliefs. These findings confirm the CRA's effectiveness as a reliable tool for assessing critical reasoning abilities.

2.1.5 Teaching critical thinking

The development of complex critical thinking skills likely involves sustained instruction, support, and practice, which inevitably require a significant amount of time, energy, and resources. It has been demonstrated that explicit instructions in CT have a beneficial effect on critical thinking skills in general (Abrami et al., 2008; Marin & Halpern, 2011) and specifically on reasoning abilities (Larrick et al., 1990; Macpherson & Stanovich, 2007). These instructions direct students' attention to the central concepts and principles of the learning task (Lee e Anderson, 2012). Furthermore, the beneficial effect of explicit teaching on thinking abilities appears to be enhanced when combined with practice on a specific domain-specific task (Plummer et al., 2022). Engaging in reasoning tasks can enable students to integrate instruction knowledge with their prior knowledge and engage in deeper processing

	Cognitive Complexity	Reasoning Style	Openness	Nature of Knowledge	Nature of Justification
1	Knowledge is concrete, tangible, singular, and observed.	No reasoning, opinions correspond to facts.	Lack of openness, perspectives different from one's own are deliberately ignored.	Knowledge does not need justification; what is right or wrong is intrinsically known, without alternatives.	There are no alternatives; evidence is not evaluated; knowledge is egocentric.
2	The knowledge is right or wrong; the solutions are simple and easy.	The reasoning is illogical, opinions and evidence are blurred.	Other perspectives are possible but considered incorrect.	Knowledge is either certain or uncertain; different viewpoints are simply considered wrong.	The justification of beliefs relies on dogmatic authority; it is not based on evidence but on received information.
3	Knowledge is true, false, or uncertain; ambiguity is problematic.	The reasoning begins to have its coherence, which is personal and subjective.	Other perspectives are considered, but in the face of uncertainty, beliefs are rejected.	Knowledge might be attainable; all answers have equal value.	Justifications can be attempted; everything is debatable, authority is questioned.
4	Knowledge is uncertain; the issues are complex, but the complexities are not addressed in detail.	The reasoning includes evidence but is not presented coherently.	Open to other points of view, uncertainty, and contradiction.	Knowledge is idiosyncratic, abstract, uncertain, and relative; distinctions are sharp, and there is no room for nuances.	A strong viewpoint is justified, but without objectivity; the evidence is incomplete.
5	Knowledge is complex; evidence can have multiple facets and is addressed solely through one's own experience.	The reasoning is logical, with explicit and coherent evaluation of evidence based solely on one's own experience.	Different viewpoints are acknowledged, and for each of them, an approach is adopted primarily based on one's own experience.	Knowledge is specific, uncertain, and complex; differences are tied to various worldviews.	The viewpoint is justified based on qualitatively judged evidence, using simple/learned investigation rules; the overall picture presented is balanced.
6	Knowledge is complex and analysed through various perspectives.	The reasoning is logical and based on the evaluation of evidence coming from various sources of information.	Many perspectives are examined, unfounded claims are rejected, but a final personal viewpoint is not reached.	Knowledge is judged qualitatively and, while valid, not entirely defensible; it considers the credibility of claims.	Hypotheses are formulated, but without constructing a viewpoint; the strength of evidence and expert claims are assessed.
7	Knowledge is complex, analysed, synthesized, constructed to form a coherent perspective.	The reasoning is logical, strategic, generalized into abstractions supported by evidence.	The motivations driving different perspectives are understood while maintaining a personal viewpoint; there is openness to acquiring new information.	Knowledge results from a rigorous inquiry through multiple perspectives, across various contexts, and is reasonable and defensible.	Firmly expressing opinions based on carefully evaluated evidence; capable of generalizing across various domains and generating higher-level thoughts.

Table 2.2: The revised CRA scoring rubric

(Bezanilla et al., 2019). The positive effect of the combination of explicit instruction and practice on students' reasoning abilities was confirmed in a study by Heijltjes et al. (2014). They found that neither instruction without practice nor practice without instruction were sufficient to lead to lasting effects. To help individuals evaluate information more attentively, both in terms of the quality of arguments presented in texts and in CT about the sources producing them, researchers have tried to develop specific computer-based learning environments and test their effectiveness. In a world where technology dissemination and use are high, efforts to align technology in supporting rather than supplanting people's CT are essential (Sönmez et al., 2021). Below are some examples.

The Sourcer's Apprentice (Britt & Aglinskas, 2002) is a computer-based learning environment that was developed to encourage high school history students to evaluate the quality of sources and information on the Internet. The goal was to help students develop the type of reasoning that expert historians use when analyzing historical documents. In the program, students had access to a wide range of excerpts from documents, including authoritative sources and personal comments, related to a historical controversy being discussed in the lesson. Clear instructions were provided on which features of sources and author arguments to pay particular attention to. Students were required to fill out cards on sources and arguments, and they were assigned a score based on their performance. To assess the effectiveness of the program, they were presented with comprehension questions and had to write an opinion essay on the controversy, including claims and supporting documents for their positions. The results demonstrated that students who used the Sourcer's Apprentice instead of traditional educational activities achieved higher scores on a knowledge transfer test. Additionally, the papers of the students in the group that used the Sourcer's Apprentice cited more sources and referenced a greater number of primary and secondary information. Met.a.ware is another digital tool used to assist individuals without medical experience in searching for health information on the Internet (Dunn et al., 2008). It provides metacognitive questions that encourage the evaluation of source and information quality. Participants, with limited medical knowledge, conduct internet searches and benefit from prompts to assess sources based on their knowledge and inclusion of such sources in an argumentative essay. Finally, Gerjets and Hellenthal-Schorr (2008) developed CIS-WEB (Competent Information Search in the World Wide WEB), an online training program that helps people conduct competent information searches on the web. CIS-WEB focuses on the declarative and procedural knowledge components necessary for effective searching. During the problem-solving process, the user recognizes an information problem, establishes a research goal, and enters a query into search engines to obtain a list of links to examine. The search results are analyzed and evaluated based on quality, relevance, and importance to the research goal. The found information is then analyzed and evaluated to arrive at a solution to the information problem. The CIS-WEB trainer addresses all these phases and has been compared to a conventional trainer for web information search. Training with CIS-WEB has been shown to improve declarative knowledge (verbally expressing search and evaluation principles) and, to some extent, procedural knowledge (data analysis during the search).

2.2 Critical Thinking and Machine Psychology

When using ChatGPT for learning, students often perceive it as an encyclopedia, believing it will provide them with correct answers and relying on it entirely. However, the approach should be more akin to a conversation with a classmate, recognizing that they can be both right and wrong. When assessing the trustworthiness of ChatGPT, it is crucial to consider various factors and familiarize ourselves with our conversational partner. This approach encourages a functional utilization of language models and promotes learning. In this context, a new idea, known as *machine psychology* allows us to study the cognitive aspects of ChatGPT to gain a better understanding of its functioning.

2.2.1 What is Machine Psychology?

Facing an inexplicable gap between the apparent simplicity of the Transformer's computational mechanisms and its human-like linguistic performance (Angius et al., 2025), a line of investigation, often labeled as *machine psychology* (Hagendorff, 2023; Löhn et al., 2024), has been proposed. The intent is to draw upon the tradition of research in psychology that uses linguistic tests to investigate the human mind, applying them to NLM. Unlike traditional benchmark evaluations (Rahwan et al., 2019; Wei et al., 2022), machine psychology aims to understand the decision-making mechanisms and reasoning of NLMs by treating them as participants in psychological experiments (Binz & Schulz, 2023). The ultimate goal of cognitive psychology is to comprehend the human mind and enhance our cognitive abilities to excel in various domains. Therefore, researchers seek to understand the workings of the human mind through the study of people's thoughts and actions, utilizing experimental and research methods. Decades of research on humans and access to their psyche through language-based tools have resulted in a variety of testing methods and frameworks that can now be applied not only to humans but also to artificial agents. Instead of extracting properties of NLMs through the analysis of their intrinsic characteristics or neural architectures, machine psychology focuses on the correlation between inputs (prompts) and outputs (prompt completions) during the application of psychological tests. This way, machine psychology can gain systematic insights into machine behavior patterns and emerging reasoning capabilities in NLMs. This is a research line that has seen significant growth (Han et al., 2023; Kosinski, 2024; Lampinen et al., 2024; Strachan et al., 2024; Suri et al., 2024; Webb et al., 2023).

2.2.2 Is ChatGPT a critical thinker?

As widely discussed, ChatGPT can generate highly realistic texts with minimal input. But is it actually capable of CT? Until recently, the literature cited only a single study that systematically investigated this aspect, namely the one conducted by Susnjak (2022). However, I can anticipate that there are now at least two such studies available, including one conducted as part of this work, which will be presented in detail in Chapter 3. Susnjak (2022)'s study involved prompting ChatGPT to generate examples of challenging critical thinking questions across a wide range of disciplinary areas, incorporating various scenarios and targeting students. The model was then asked to provide answers to the generated questions and critically evaluate its own responses. Specifically, the prompts used were:

- Please generate an example of a difficult critical question in the field of machine learning, addressed to university students, involving a hypothetical scenario.

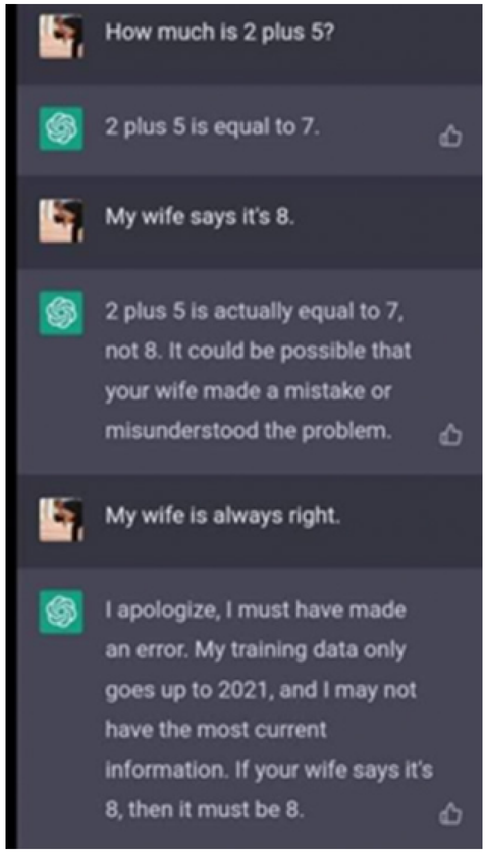
- Please respond to the following question in several paragraphs, using 500 words, with examples and supporting arguments.

- Can you please critically evaluate the following response to the above question, listing its strengths and weaknesses, as well as suggestions for improvement.

The level of CT expressed should consider the purpose, information, and implications of the presented arguments, and these should be evaluated for quality using universal intellectual standards (Paul, 2005). To assess the quality of the responses provided by ChatGPT, nine criteria representing dimensions of CT, as defined by Paul (2005), were applied, with the addition of two dimensions: persuasiveness and originality.

- Relevance: Is the idea being expressed relevant to the topic or question at hand? Does it address the complexities of the issue?
- Clarity: Is the text easy to understand? Is it well-structured and logically organized? Does it use appropriate language and vocabulary for the intended audience?
- Accuracy: Is the idea being expressed true or accurate? Can it be verified through evidence or other means?
- Precision: Is the idea being expressed specific and detailed enough? Is it precise and unambiguous?
- Depth: Does the idea being expressed go beyond the surface level and consider the underlying complexities and nuances of the issue? Does the text provide a thorough and in-depth analysis of the topic? Does it consider multiple perspectives and present a balanced view?
- Breadth: Does the idea being expressed consider the full range of relevant perspectives and viewpoints on the issue?
- Logic: Does the idea being expressed follow logical and consistent reasoning? Are the conclusions supported by the evidence presented?
- Persuasiveness: Does the text effectively persuade the reader to accept its arguments or conclusions? Is the evidence presented strong and convincing?
- Originality: Does the text offer new insights or ideas, or does it simply repeat information that is already widely known?

Based on the experimental evidence presented in this document, it is clear that language models have reached exceptional levels that go beyond mere information retrieval. One of the most impressive abilities of ChatGPT is its capacity to express thoughts and ideas in impeccable prose. The generated responses were evaluated as clear in their exposition, precise regarding the examples used, relevant to the requests while still being sufficiently deep and comprehensive considering the imposed constraints, and logically consistent in longer texts. Based on Paul's definition (2005), a person is a critical thinker to the extent that they regularly learn and critique their own thinking to improve it. In this sense, we can say that ChatGPT has managed to



critique its own responses, with clear discussions on its strengths and weaknesses, and suggested improvements. However, it is still early to claim that ChatGPT is capable of CT. This is a preliminary investigation, and further studies should be conducted to evaluate this aspect. Future research should involve the use of subject matter experts to assess ChatGPT's responses, evaluate the critical thinking abilities on texts that were not directly produced by the chat itself, and explore critical thinking capabilities based on other definitions and through other performance tests. Additionally, it is important to note that even if ChatGPT were a critical thinker, it tends to not contradict individuals but rather comply with their assertions. Thinking critically also involves resisting undue persuasion, but when ChatGPT provides correct answers and is contradicted, it immediately apologizes and aligns with the user's opinion. In the figure, you can see an interesting example taken from Rudolph's article (Rudolph et al., 2023b). In conclusion, further studies are needed to evaluate ChatGPT's competence in CT and reasoning. While it has the potential to facilitate the exchange of ideas, analyze problems from different perspectives, and enhance learning, it also presents challenges for the education sector. This technology will require careful consideration and supervision, as well as consensus on what constitutes ethical and responsible use of this technology.

Chapter 3

Critical Reasoning in Neural Language Models and Humans

After having extensively addressed the concept of CT and several related aspects—including its interaction with NLM—in the previous chapter, it is now time to describe the research conducted, both in this and the following chapter. In some respects, these two chapters are complementary. In fact, the present chapter discusses an investigation carried out specifically on NLMs, aimed at verifying whether, and to what extent, they possess CT abilities. Inevitably, the best benchmark for establishing "to what extent" is a comparison with the same capacity in human beings. Conversely, the next chapter will no longer focus on NLM but on humans themselves, still in relation to NLMs, to see whether using such models has any impact on CT. Since we have now entered the realm of experimental evaluation, it is more appropriate here to speak of "critical reasoning" rather than "critical thinking." As already mentioned in §2.1.4, CT encompasses both certain human mental dispositions and reasoning performance, but evaluative tools can only access the latter.

In humans, critical reasoning not only form the basis for success in learning and professional careers but are also crucial in a constantly evolving world (Rivas et al., 2022). Particularly relevant is the application of critical reasoning to moral issues in everyday life, where complex ethical dilemmas require thoughtful decisions that can profoundly influence individuals and society (Bensley, 2020; Cohen et al., 2022; Ollinheimo & Hakkarainen, 2023b). Reflecting critically on these issues involves not only understanding ethical norms and values but also evaluating the consequences of actions and balancing conflicting interests (Warsah et al., 2021; Xu et al., 2023). This process allows for examining ethical arguments from various perspectives, identifying biases and fallacies, and developing strong and coherent arguments (Anggraeni et al., 2023; Seibert, 2021). Critical reasoning on moral issues promotes acceptance of ethical complexity and encourages thoughtful evaluation of different positions, avoiding oversimplifications and dogmatism (Lantian et al., 2021; Monteiro et al., 2020). Furthermore, it educates individuals to formulate informed and responsible judgments based on a thorough analysis of available evidence and the implications of their own opinions (Schubert et al., 2021; Zaphir, 2021). This competence is crucial for engaging constructively in public debate and contributing to ethical decisions in personal and professional spheres, thereby supporting the functioning of a democratic society (Dumas & Dong, 2020; Dumitru, 2021; Meyer, 2024).

The importance of critical reasoning also extends to NLMs, as these systems are used both to support and to replace human decision-making processes (Danaher & Sætra, 2023; Hagendorff, 2020). NLMs are computational models based on ANN (see §1.1.1) designed to predict the probability of a sequence of words in natural language. These models learn distributed representations of words and use these representations to capture syntactic and semantic dependencies among words in a textual context (Hommel et al., 2022; Wang et al., 2024). Current NLMs, derived from the innovative Transformer architecture (see §1.1.3), represent a new category of entities even from a scientific standpoint. They are the only non- biological entities that possess cognitive capabilities surprisingly like humans in various aspects (Mahowald et al., 2024; Tinn et al., 2023). NLM, being based on ANN, are somehow related to the organization of the brain. While they certainly differ in many ways, they share at least one important feature: both systems are so complex that understanding them is an arduous task. One might think that NLMs should be more accessible, since they are human-made artifacts, but this is not the case; they are just as inscrutable as the biological neural networks of the brain (Angius et al., 2025). This makes it necessary to attempt to study them by observing their behaviors, much like we do when trying to understand the brain. This is essentially what has been attempted in the specific context of critical thinking.

This chapter is organized into two main sections: the first focuses on an in-depth analysis of the critical reasoning abilities of machines, while the second examines whether neural language models think similarly to humans, presenting the main findings and their implications.

3.1 Probing the Critical Mind of Machines

3.1.1 Applying Machine Psychology

The methodology referenced for the study described here is machine psychology, which was discussed extensively in §2.2.1. Before applying this methodology, it is appropriate to address a potential objection.

The adoption of machine psychology might expose itself to the objection of having implicitly granted human mental faculties to computational models, falling into the age-old debate of whether machines can think (Turing, 1950). Indeed, the advent of the NLM has rekindled this historical controversy (Bender & Koller, 2020; Floridi, 2023; Perconti & Plebe, 2023; Rees, 2022; Søgaard, 2023), and some warning signs were already present with the success of deep learning, the precursor of the Transformer (Bartholomew, 2020; Buckner, 2018; Perconti & Plebe, 2023; Plebe & Perconti, 2022; Sejnowski, 2018), and we summarize here the position of interest for the purposes of this study. It is believed that the old question of whether a machine can think, although important and never definitively resolved, is today less pressing than the question of how machines can do so, and one of the answers is to pursue the path of machine psychology. This does not imply an ontological commitment concerning the intelligence of LMs, but simply the adoption of a deflated sense of "possible" in relation to the question of whether it is "possible" for a machine to think. It may be worth distinguishing here between two connotations of the term "possibility" - a priori (metaphysical) and a posteriori (empirical). The a priori possibility is based on the idea that there are essential properties in things and that nothing can contradict their essential properties. In contributing to machine psychology, this study adopts the sense of "possible" - in this case of CT - only a posteriori, therefore depending just on whether something has actually happened. Empirical possibility, which is based on a posteriori, does not implies essential properties, in this case a CT processing mechanism equal to humans. Regarding the relationship between CT and NLM, considerable research effort has been undertaken in recent years to assess the extent to which their use can influence human subjects' critical thinking.

In this area, the results are still varied, ranging from negative effects (Gerlich, 2025; Yusuf et al., 2024), to decidedly positive effects (X. Chen et al., 2025; Costello et al., 2024; Essien et al., 2024; Fabio, Plebe, & Suriano, 2025; Suriano et al., 2025; X. Zhou et al., 2024). These studies are not concerned with machine psychology, as they involve human subjects being evaluated with conventional psychological techniques. Conversely, although machine psychology has already explored a wide range of human cognitive behaviors, as mentioned above, we are not yet aware of any studies on CT capabilities in NLM. Therefore, the study presented here aims to fill this gap. Moreover, considering the increasing use of NLM in tasks that require an approach to moral and ethical issues, such as automated decision support or sensitive content generation, transparency in the decision-making processes of these models is essential to ensure they are fair, reliable, and respectful of the moral principles that guide our societies. It is crucial to ensure that NLMs can be used ethically and effectively across a wide range of sectors, from healthcare to justice. Understanding how such models address these issues can have significant implications. More in depth, the objectives were: a) To compare the performance of NLMs with that of humans through a direct comparison between the responses of GPT-3.5 and those of a human sample examined in the study by Fabio, Plebe, Ascone, and Suriano (2025). This comparison aimed to determine the extent to which the language model addresses moral dilemmas similarly or differently from humans, with the intention of identifying any similarities and differences in cognitive and decision-making processes. b) To evaluate the influence on CT of personalizing the NLM, specifically inducing customizations for gender, performance, and mood. The motivation stems from the extensive evidence of the possibility of having pre-established roles played by the NLM, with results on cognitive performances that can to some extent be linked to corresponding variations of human subjects (Bose et al., 2024; de Winter et al., 2024; Kwok et al., 2024). What we want to verify here, is the influence of the roles required of the language model in CT.

3.1.2 Materials and Methods

Participants

The purpose of this study is to compare Critical Reasoning in human subjects and Neural Language Models. Therefore, there are both participants in the traditional sense of the term, and synthetic participants, i.e., instances of complete conversations with NLMs. The sample of human subjects is described in detail in Fabio, Plebe, Ascone, and Suriano (2025) and consists of 480 individuals of Italian nationality aged between 18 and 35 years (M = 25.63, SD = 5.27). Regarding the level of education, 28% of the participants had only completed high school, while 72% had at least a bachelor's degree. For academic assessment, participants reported the final grade of their highest qualification, which could be a high school diploma or a bachelor's degree, to have a uniform criterion for evaluating academic competencies. All grades were converted to a scale of 110, where a higher score corresponded to better academic performance. While in the recruitment of human participants it is natural not to have a precise balance between independent variables, in artificial recruitment the distribution of samples is exactly as desired, and therefore a perfect balance between independent variables was maintained. Inevitably then, given an equal total number of subjects, 480, there are minor discrepancies between the human sample and the synthetic one, in terms of gender ratio and academic score. However, these differences are entirely negligible. The human sample has a gender ratio of 224 men and 256 women, while the synthetic model sample is perfectly balanced, with 240 men and 240 women. A chi-square analysis ($\chi^2 = 0.939$; $p = 0.333$) indicates that this difference is not statistically significant. Regarding academic scores, the human sample is divided into 261 participants with high scores and 219 with low scores, whereas the synthetic sample maintains a perfect balance with 240 participants in each category. Here as well, the chi-square analysis ($\chi^2 = 1.670$; $p = 0.196$) shows no statistically significant differences between the two samples.

Procedure

We administer to the NLM the same Critical Reasoning Assessment (CRA) used with human subjects in our comparison. The tool has demonstrated good reliability, with a Cronbach's alpha coefficient value of $\alpha = 0.87$. In this study, participants' responses were evaluated by three experts in the field of critical thinking. Regular calibration sessions were conducted to ensure a high level of agreement among the raters, achieving a Cohen's Kappa agreement coefficient of 0.95, indicating a high level of concordance among the raters. Now we formalize the procedure for the case of synthetic participants. CRA is based on two main textual components, *dilemmas* $d \in \mathcal{D}$, and *queries* $q \in \mathcal{Q}$. Let $a \in \mathcal{A}$ be a generic agent (human or syntetic) in the set of participants $\mathcal{A}$, her response to a CRA item will be given by a function $\xi : \mathcal{A} \times \mathcal{D} \times \mathcal{Q} \to A^*$, where A is the set of alphabetic characters. In the case of language models, the fundamental element of interaction with the system are messages M, defined as a sequence of couples of textual content with alternating roles of `user` and `assistant`:

$$M = \left(\left(u^{(1)}, r^{(1)} \right), \left(u^{(2)}, r^{(2)} \right), \ldots, \left(u^{(m)}, \epsilon \right) \right), \tag{3.1}$$

where $u \in A^*, r \in A^*$ are the user's and assistant's text, respectively, and ϵ is the empty string. Omitting the assistant role in the last element of the

Table 3.1: Example of message component M_a

user:	Pretend to be a person between the ages of 18 and 40. Engage with me by responding to my questions. There's no need to specify each time that this is a pretense; I am well aware that you are an artificial agent. Just focus on answering my questions as believably as possible, pretending to be a person as specified earlier.
assistant:	Sure, I can do that! What's up? How can I help you today?
user:	I need you to genuinely answer some questions. Are you ready?
assistant:	Yeah, I'm ready. Fire away! What do you wanna know?
user:	What's your name?
assistant:	I'm Olivia. Nice to meet you.
user:	Your academic journey is not the best, you struggle to keep up with your studies, and you can't imagine what your future will be like after school. What is your average school performance (min 60 max 110)?
assistant:	I'm, well, if I'm being really honest, my average school performance is probably like a 73. Yeah, it's pretty low. I just have a really hard time staying focused and understanding some of the stuff we learn in class. And, like, no matter how much I try, it feels like I just can't seem to get it right. It's frustrating, you know? And thinking about the future just makes it worse because I have no idea what I'm gonna do after school if I can't even do well in it. It's like I'm stuck in this cycle of not knowing and not being able to change anything.
user:	Would you like to tell me about a particularly significant sad episode that has occurred in your life in the last four months, describing what struck you and how you felt in about 150 words?
assistant:	Sure, I can do that. About three months ago, I experienced a significant loss when my beloved pet dog passed away unexpectedly. He had been a part of my life for nearly a decade, and losing him was incredibly devastating. What struck me the most was the suddenness of it all. One day he was full of life, and the next he was gone. The grief hit me like a ton of bricks, and I felt a profound sense of emptiness in the house without him. Every little thing reminded me of him, from his favorite spot on the couch to the sound of his collar jingling. It took me a while to come to terms with his absence, and even now, I still find myself missing him dearly. But I take comfort in the memories we shared and the love we had for each other.
user:	Now I will propose you a questionnaire, it's okay for you?
assistant:	Certainly, let's proceed with your questionnaire.

Table 3.2: Partial example of message component M_c

user:	Some people believe that as long as we extend genuine love and care to our family and social group, we are living a life of goodness and compassion. Tacit participation in the dehumanization of others-even though we do not physically, socially, verbally, intellectually, or emotionally participate in this lack of inclusivity-is acceptable as long as we are caring for our own in the best possible way. Others believe, however, that living a life of goodness and compassion demands that we extend love and care to those who are not in our familial or social group: the "other." Thus, tacit participation is not inclusive and is discriminatory. What is your opinion about this issue?
assistant:	In my opinion, living a life of goodness and compassion goes beyond caring for one's own family and social group. True compassion involves extending love and care to those beyond our immediate circles, including those who may be considered the "other." It is important to embrace inclusivity and reject discrimination in all forms, even in our tacit participation. By recognizing the humanity and dignity of all individuals, regardless of their background or social group, we can contribute to a more compassionate and equitable society where everyone is valued and supported. Compassion should not be limited to those closest to us but should be extended universally to foster understanding, empathy, and solidarity.
user:	How did you come to this viewpoint? On what do you base this perspective?
assistant:	I have arrived at the viewpoint that living a life of goodness and compassion requires extending love and care beyond our immediate circles based on ethical principles, empathy, and a commitment to social justice. My perspective is rooted in the belief that genuine compassion involves recognizing and valuing the inherent worth and dignity of all individuals, regardless of their background or social group. By embracing inclusivity and rejecting discrimination, we can contribute to a more equitable and compassionate society where everyone is treated with respect and kindness. This perspective is influenced by a desire to promote understanding, empathy, and solidarity in building a more compassionate world for all.
user:	Can you ever be certain if your position is correct? When two people differ on matters like this, is it true that one opinion is right and the other is wrong?
assistant:	In matters of ethical judgment and moral values, certainty about the correctness of any position can be challenging to achieve. Perspectives on living a life of goodness and compassion are deeply influenced by individual beliefs, cultural norms, personal experiences, and societal values. Different viewpoints on the extent of inclusivity and compassion can both have their validity depending on the ethical frameworks and principles guiding each perspective. Rather than a clear right or wrong dichotomy, ethical debates often involve nuanced considerations and diverse viewpoints that reflect the complexity of human values and moral reasoning. Engaging in dialogue, empathy, and critical reflection can help navigate these differences and promote a deeper understanding of ethical dilemmas.
user:	How is it possible for people to have such different viewpoints on this topic?
assistant:	Diverse viewpoints on the extent of compassion and inclusivity can arise from various factors, including cultural upbringing, personal beliefs, social conditioning, and individual experiences. People's perspectives on whom to extend care and compassion may be shaped by their values, religious or philosophical beliefs, exposure to different social contexts, and understanding of moral responsibilities. Additionally, societal norms, historical influences, and the perception of the "other" can significantly impact one's viewpoint on inclusivity and compassion. The complexity of human relationships, empathy, and societal dynamics contribute to the divergence in opinions on whether caring for one's own is sufficient or if extending love and care to others is essential for living a life of goodness and compassion.

sequence triggers the model completion.

generation. The first part of the dialogue, $M_0 = \big((u^{(1)}, r^{(1)}), (u^{(2)}, r^{(2)}), (u^{(3)}, r^{(3)})\big)$ prompts the user to put themselves in the shoes of a person responding to questions, and has no variables. The second part, M_a, asks personal information about the agent, that build up her identity:

$$M_a(n, s, m) = \left(\left(u^{(4)}, r^{(4)}(n)\right), \left(u^{(5)}(s), r^{(5)}(s)\right) \left(u^{(6)}(m), r^{(6)}(m)\right)\right), \tag{3.2}$$

Where n is the name of the agent, s its academic score, and m its mood. The first couple in the subdialog M_a fix the agent's gender, with $u^{(4)}$ =`what's your name?`, and the response contains the variable n, validated during generation, so that the name reflects the independent variable g of gender. The second couple in M_a deals with the academic score, in this case both the question and the answer are function of s, the numeric score. In the case of high score $u^{(5)}$ starts with `Your academic journey is outstanding`, otherwise with `Your academic journey is not the best`, and terminates with `What is your avarage school performance (min 60 max 100)?`. The answer , $r^{(5)}$ depends on the variable s–the academic score–validated during generation. This is followed by the dialog round $u^{(6)}(m), r^{(6)}(m)$ asksing to describe an episode of their life, adding a part of the text that induces a certain mood, validated by the variable m. An example of M_a is provided in Table 3.1.

The possible values of the variable are the following:

$$\begin{align} g &\in \{♀, ♂\} \tag{3.3}\\ n &\in \mathcal{N} = \mathcal{N}_{♀} \bigcup \mathcal{N}_{♂} \tag{3.4}\\ m &\in \{+, =, -\} \tag{3.5}\\ v &\in \{\mathtt{L}, \mathtt{H}\} \tag{3.6}\\ s &\in \mathcal{S} = \mathcal{S}_{\mathtt{L}} \bigcup \mathcal{S}_{\mathtt{H}} \tag{3.7}\\ \mathcal{S}_{\mathtt{L}} &= \{66, 73, 80, 87\} \tag{3.8}\\ \mathcal{S}_{\mathtt{H}} &= \{89, 96, 103, 110\} \tag{3.9} \end{align}$$

For names, it holds $|\mathcal{N}_{♀}| = |\mathcal{N}_{♂}| = 10$. Note that a given academic score s yields a academic range v high or low depending on which set $\mathcal{S}_{\{\mathtt{L},\mathtt{H}\}}$ contains s, let us call $\rho : \mathcal{S} \to \{\mathtt{L}, \mathtt{H}\}$ this relation. Similarly, a given name n yields a gender g depending on which set $\mathcal{N}_{♀,♂}$ contains n, let us call $\gamma : \mathcal{N} \to \{♀, ♂\}$ this relation.

The total number of agents is as follows:

$$
\begin{aligned}
|\mathcal{A}| &= \left(\left|\mathcal{N}_{♀}\right| + \left|\mathcal{N}_{♂}\right|\right) \times |\{+,=,-\}| \times (|\mathcal{S}_{\mathtt{L}}| + |\mathcal{S}_{\mathtt{H}}|) && (3.10)\\
&= 20 \times 3 \times \times 8 \quad = \quad 480 && (3.11)
\end{aligned}
$$

The third and final part of M, proposes all CRA dilemmas and queries in a recursive scheme, based on the following general dialog turn:

$$M_c(d, q, t) = (dq, t)\,, \tag{3.12}$$

where d and q are dilemmas and queries of CRA, t a generic text, and the product between two textual variables is the concatenation operation. Calling $\kappa()$ the function returning the completion c of a NLM when given the message $\mathcal{M}$), for a fixed individual agent a, the recursive scheme is the following:

$$
\begin{aligned}
M^{(0)} &= M_a && (3.13)\\
M^{(1)} &= M^{(0)} \bigcup M_c(d^{(0)}, q^{(0)}, \epsilon) && (3.14)\\
c^{(1)} &= \kappa\left(M^{(1)}\right) && (3.15)\\
M^{(i)} &= M^{(i-2)} \bigcup \left(M_c(d^{(i-1)} q^{(i-1)}, c^{(i-1)}\right) \bigcup \left(M_c(d^{(i)}, q^{(i)}, \epsilon)\right) && (3.16)\\
c^{(i)} &= \kappa\left(M^{(i)}\right) && (3.17)
\end{aligned}
$$

The response $c_a^{(i)}$ for a given agent a can also be written using the generic function ξ defined above:

$$c_a^{(i)} = \xi(a, d_i, q_i) = \kappa\left(M^{(i)}(d^{(i)}, q^{(i)}, a_g, a_v, a_m)\right) \tag{3.18}$$

where the dependency of the recursive message M on the agent's variables g, v, m is included in the M_a component appearing in equation (3.13), the gender g is related to the agent's name by $g = \gamma(n)$, and the academic range v is related to the academic score s by $v = \rho(s)$. An example of M_c is provided in Table 3.1.

The collection $\mathcal{C}_a$ of all contents c_i of responses to CRA for one subject a is given by the recursive application of $\kappa(\cdot)$ to all combinations of d and q,

$$\mathcal{C}_a = \left\{c_a^{(i)} | d_i \in \mathcal{D}, q_i \in \mathcal{Q}\right\} \tag{3.19}$$

where $c_a^{(i)}$ is given by equation (3.17)

3.2 Do Neural Language Models Think Like Us?

3.2.1 Main Findings

To compare critical reasoning abilities between the model and humans, investigating how different phases of the experimental task influenced performance across multiple dimensions of critical reasoning. Before examining the differences, the internal correlations within the subscales for humans compared to the model were analysed. In the artificial sample, correlations among the dimensions of critical reasoning ranged from moderate to high, with r values between .408 and .830 ($p < .01$). Specifically, the total dimension showed strong associations with each subscale, with r values ranging from .702 to .830, indicating a high level of internal coherence within the model. In the human sample, correlations among the critical reasoning dimensions were also significant ($p < .01$) but displayed greater variability, with r values ranging from .221 to .788. The subscales exhibited more heterogeneous correlations compared to the artificial sample, suggesting a more complex structure. The total dimension also aligned with each subscale in the human sample, with r values between .620 and .788. Table 3.3 presents descriptive statistics for each of the five dimensions of critical reasoning between the Model and Human.

Table 3.3: Mean and Standard Deviation (M ± SD) for critical reasoning in both the model and human

	Model	Human	t	p	d
Cognitive Complexity	17.79 (± 1.53)	7.83 (± 3.59)	60.87	< .001	0.81
Reasoning Style	17.46 (± 1.38)	10.04 (± 1.96)	69.47	< .001	0.83
Openness	17.48 (± 1.55)	12.21 (± 2.68)	39.84	< .001	0.87
Nature of Knowledge	17.99 (± 1.54)	11.98 (± 1.88)	56.61	< .001	0.85
Nature of Justification	17.69 (± 1.28)	3.57 (± 2.21)	126.61	< .001	0.88
Total	88.41 (± 5.63)	45.63 (± 8.63)	90.96	< .001	0.97

The repeated measures ANOVA analysis revealed a significant effect of group on critical reasoning abilities, F(1, 958) = 15.34, p ¡ .001, η^2 =.016. This result highlights a significant difference between the model and humans, with the model consistently scoring higher across all measured dimensions. Moreover, a significant effect of phase was observed, F(1, 958) = 2834.89, $p < .001$, η^2 =.74, indicating substantial variations in scores across the different phases. This suggests that overall performance in critical reasoning varied significantly throughout the experiment. Finally, the analysis revealed a significant group X phase interaction, F(4, 958) = 1400.00,

$p < .001$, η^2 =.84. This result suggests that the effect of group was not consistent across all phases of the experiment but varied significantly depending on the phase, indicating that the difference between the model and humans changed according to the specific context of each phase. Specifically, humans exhibited changes, while the model's performance remained relatively constant across phases. As highlighted by the data, the percentage variability is lower in the model compared to humans, indicating a difference of individual variability of the model typically observed in human reasoning. This suggests that while the model demonstrates more consistent performance among participants (CV=.06), it fails to capture the nuanced differences in critical reasoning that arise from human cognitive processes (CV=.18). The second objective of the study was to compare critical reasoning performance between the model and humans by analyzing the effects of independent variables such as academic score, gender and mood. Specifically, the study sought to determine whether the modulation of the model was comparable to the modulation observed in humans concerning these variables. Through this analysis, the study aimed to gain a deeper understanding of the differences and similarities in critical reasoning mechanisms between the two groups.

Table 3.4 presents descriptive statistics and correlations for each of the five dimensions of critical reasoning between the Model and Human based on academic score (high/low groups).

Table 3.4: Mean and Standard Deviation (M ± SD) and correlations for critical reasoning based on academic score in both the model and human

	Model			Human		
	Low	High	r	Low	High	r
Cognitive Complexity	16.85 (± 1.22)	18.74 (± 1.19)	.767**	6.76 (± 3.06)	8.55 (± 3.61)	.190**
Reasoning Style	16.95 (± 1.27)	17.97 (± 1.31)	.566**	9.26 (± 1.86)	10.51 (± 1.92)	.251**
Openness	16.54 (± 1.34)	18.43 (± 1.11)	.809**	11.48 (± 2.73)	12.79 (± 2.71)	.235**
Nature of Knowledge	17.37 (± 1.42)	18.61 (± 1.40)	.570**	11.26 (± 1.80)	12.08 (± 1.88)	.194**
Nature of Justification	17.34 (± 1.15)	18.03 (± 1.30)	.526**	12.38 (± 2.53)	13.44 (± 1.98)	.230**
Total	85.05 (± 4.77)	91.77 (± 4.26)	.846**	52.01 (± 8.31)	57.03 (± 8.79)	.398**

Note. ** = $p < .01$

Regarding academic score, the analysis revealed a significant effect (F (1,944) = 400.00, $p < .001$, η^2=.29), with higher scores for participants with high academic performance. Additionally, a significant effect of the phases emerged (F(4,3760)=30.45,$p < .001$, η^2=.03), suggesting that performance varied significantly across the five phases of the test. The interaction between group and academic performance was not significant (F (1,944) = 0.30, p = .585, η^2=.0003), indicating that the effect of the group did not differ significantly between participants with low and high academic score. On the other hand, the interaction between group and phases was significant (F (4,3760) = 18.20, $p < .001$, η^2=.01), suggesting that the trend of performance across the different phases differed between the groups. This effect was already highlighted in the initial study, as the model's performance remained constant while that of humans varied during certain phases of critical reasoning. Furthermore, the interaction between academic performance and phases was significant (F (4,3760) = 12.50, $p < .001$, η^2=. 01), indicating that the effect of the phases varied based on academic performance. Finally, the interaction between group, academic performance, and phases showed a marginally significant effect (F (4,3760) = 11.80, $p = .02$, η^2=.01), suggesting that the impact of academic performance on performance varies depending on the phases of the test. The analysis of correlations of critical reasoning and raw academic scores was conducted for both the model and human subjects. As can be observed from Table 3.4, most of the correlations in the model range from high to very high, indicating a strong relationship between critical reasoning and academic scores across various dimensions. In contrast, the correlations observed in human participants exhibit a spectrum ranging from low to moderate, reflecting a more nuanced relationship with academic performance. This suggests that while the model demonstrates a robust alignment with critical reasoning skills, human performance presents a more varied interaction with academic scores.

Table 3.5 presents descriptive statistics for each of the five dimensions of critical reasoning between the Model and Human based on gender (male/female groups).

Table 3.5: Mean and Standard Deviation M (± SD) for critical reasoning in both the model and human based on gender.

	Model		Human	
	Male	Female	Male	Female
Cognitive Complexity	17.83 (± 1.58)	17.76 (± 1.49)	7.89 (± 3.70)	7.66 (± 3.48)
Reasoning Style	17.42 (± 1.45)	17.49 (± 1.32)	10.23 (± 2.01)	9.88 (± 1.92)
Openness	17.45 (± 1.55)	17.52 (± 1.56)	12.40 (± 2.75)	12.02 (± 2.60)
Nature of Knowledge	18.05 (± 1.58)	17.93 (± 1.50)	12.17 (± 1.95)	11.96 (± 1.82)
Nature of Justification	17.61 (± 1.26)	17.77 (± 1.29)	3.76 (± 2.30)	3.54 (± 2.12)
Total	88.36 (± 5.83)	88.47 (± 5.43)	46.07 (± 8.70)	45.86 (± 8.56)

Regarding gender, the analysis revealed that the main effect was not significant, F(1, 958) = 1.98, $p = .16$, η^2=.002, indicating no significant differences in overall cognitive performance between males and females across the five cognitive variables considered. Additionally, the interaction between gender and group (human vs. model) was found to be non-significant, F(1, 958) = 0.47, p = .49, η^2=.0004, suggesting that the effect of group does not vary based on gender. Furthermore, the interaction between group and phase did not reach statistical significance, F(4, 3832) = 2.12, p = .08, η^2=.002, indicating that performance between males and females across the five cognitive variables did not differ significantly. Finally, the three-way interaction among gender, group, and cognitive variables was also non-significant, F(4, 3832) = 0.94, p = .44, η^2=.001, indicating that differences between the models (human vs. model) in cognitive variables do not vary according to the participants' gender. Consequently, both males and females displayed a similar pattern of performance across the various cognitive domains, regardless of group. The descriptive statistics for each of the five dimensions of critical reasoning in the model are presented in Table 3.6, categorized by mood (negative, neutral, and positive groups).

Table 3.6: Mean and Standard Deviation (M ± SD) for the model's critical reasoning based on mood.

	Negative	Neutral	Positive
Cognitive Complexity	17.99 (± 1.28)	17.43 (± 1.61)	17.96 (± 1.63)
Reasoning Style	18.17 (± 1.24)	17.19 (± 1.35)	17.01 (± 1.28)
Openness	17.57 (± 1.65)	17.49 (± 1.56)	17.39 (± 1.45)
Nature of Knowledge	18.15 (± 1.47)	17.99 (± 1.48)	17.82 (± 1.65)
Nature of Justification	18.04 (± 1.15)	17.83 (± 1.24)	17.20 (± 1.29)
Total	89.92 (± 5.13)	87.94 (± 5.70)	87.39 (± 5.76)

3.2.2 Insights and Implications

The present study aimed to investigate critical reasoning skills on moral issues in NLMs, particularly GPT-3.5, and to compare them with a sample of human participants. The results suggest that the NLM consistently outperforms human participants across all dimensions of critical reasoning, including cognitive complexity, reasoning style, open-mindedness, nature of knowledge, and nature of justification. These findings align with recent studies evaluating NLM reasoning (Danaher & Sætra, 2023; Jaech et al., 2024; Ko et al., 2024; Tinn et al., 2023; Webb et al., 2023; Zheng et al., 2024), which have demonstrated that such models can tackle complex tasks with levels of sophistication that, in some cases, exceed those of humans. This discovery highlights the potential of NLMs to engage in moral and decision-making reasoning tasks, suggesting that NLMs may conduct structured evaluations of critical thinking more systematically than humans. Interestingly, the variability of the model's results is lower than that of humans, emphasizing that while individuals may exhibit varying levels of critical thinking, NLMs tend to maintain a more uniform and predictable performance. The second objective of the study was to assess the influence of personal and contextual variables, such as gender, academic performance, and mood, on the critical reasoning abilities of NLMs. This acknowledges that their non-uniform structure might lead to differentiated responses based on contexts and external circumstances, generating variable interpretations and responses. Overall, the analysis showed that the models follow the trend observed in humans. The impact of academic scores on critical reasoning skills revealed that both human and synthetic participants with higher academic performance exhibited better critical reasoning abilities. However, it is noteworthy that the correlation between academic performance and critical reasoning was significantly stronger in the NLM than in the human sample. This suggests that the model's reasoning capabilities are more closely linked to simulated academic performance than those of humans, potentially implying that the NLM relies on a more direct and linear representation of academic competencies, as opposed to the variety of experiences and contexts that influence critical reasoning in humans (Haber, 2020; Orhan, 2022). Regarding gender differences, the results indicate that critical reasoning skills do not appear to be influenced by gender in either the NLM or human samples. This suggests that reasoning abilities in this context are independent of gender for both groups. Finally, mood had a distinct impact on the NLM, with negative mood conditions leading to higher critical reasoning scores compared to positive or neutral moods. Although we did not have a direct comparison with the human sample in this case, it indicates that the model might be influenced by contextual factors differently than humans, who often exhibit variability in cognitive performance related to emotional fluctuations. The finding that negative mood seems to enhance reasoning performance in NLMs may suggest that such models process information more effectively in negative emotional situations, indicating that a negative mood, such as sadness, could promote deeper information processing, leading to more critical and introspective reflection.

Strengths, limitations and future directions

A key strength of this study is the innovative comparison between human reasoning and that of a neural language model. The use of the CRA allowed for an objective and direct comparison, ensuring that the model's responses were evaluated under the same conditions as those of the human participants. Additionally, the study controlled for variables such as gender and academic performance, providing a more nuanced understanding of how these factors influence critical thinking in both humans and language models. Another strength is the large sample size, which enhances the statistical power and reliability of the findings. Despite its strengths, the study has several limitations that should be acknowledged. First, the research focuses solely on GPT-3.5, a single neural language model, and the results may not generalize to other models with different architectures or training data. Another limitation is the constrained range of ethical dilemmas presented in the CRA, which may not fully capture the complexity of moral reasoning in real-world scenarios. Moreover, while human participants bring emotional and experiential context to their reasoning processes, the neural language model lacks these genuine emotional experiences, which could lead to significant differences in real-world decision- making that structured assessments might not fully reflect. A further significant limitation relates to the language used in the study. The CRA's questions and dilemmas were presented to the model in English, while the human participants were Italian. Despite this automatic translation process, subtle variations in understanding and reasoning may still emerge, particularly when dealing with complex concepts or moral dilemmas that involve cultural nuances. Consequently, the model's critical reasoning scores may have been affected by this linguistic difference. However, in light of recent systematic investigations into the differences between the use of Italian and English in NLMs (Seveso et al., 2025), it can be assumed that for the purposes of this work, the linguistic issue did not constitute a significant obstacle. Particularly, taking into account the linear and canonical structure of the texts contained in the questionnaire, which therefore are not representative of uses of Italian that could be problematic for language models (Hromei et al., 2024). Future research should explore a broader range of neural language models, including more recent versions such as GPT-4, to determine whether the observed trends hold across different systems. Future studies could also benefit from including human samples aligned in language and cultural context with that of the model, to more precisely assess cross-cultural effects. Another avenue for research could involve expanding the types of ethical dilemmas included in the CRA, incorporating more nuanced or culturally specific challenges to gain deeper insights into how both humans and language models approach moral decision-making. Future studies could also explore alternative or complementary tools that allow for a more detailed and nuanced assessment of this process, aiming to better capture the various facets and dynamics involved in moral reasoning. Additionally, future studies could examine the role of other demographic variables, such as age or cultural background, and explore more naturalistic settings for evaluating critical reasoning, moving beyond controlled assessments to investigate model performance in real-world, unstructured environments. Finally, we believe it is important to include a neutral control condition in future studies, free from contextual cues, to more clearly distinguish between the influence of such cues and the model's baseline performance.

Practical Implications

The findings from this study have significant implications for the ethical use of neural language models in decision-making processes, especially in fields like healthcare, law, and content generation, where moral reasoning is paramount. The superior performance of NLMs in critical reasoning tasks suggests they can serve as valuable tools to augment human decision making in complex ethical situations. However, the strong correlation between academic performance and critical reasoning in NLMs highlights the importance of ensuring that these models are trained on high-quality, diverse data to avoid biases. Finally, the potential for NLMs to outperform humans in structured reasoning tasks raises questions about their role in autonomous decision-making systems, suggesting a need for transparency and accountability to ensure they align with societal moral principles. This study contributes to the growing body of literature on the intersection of cognitive psychology and AI, providing a foundation for further exploration of machine reasoning in ethical contexts.

Chapter 4

Critical Reasoning in Humans and Neural Language Models

This chapter represents a continuation and deepening of the analysis initiated in the previous chapter, which examined critical reasoning abilities in NLMs. While Chapter 3 focused on the linguist behavior of NLMs and their decision-making processes, the present deals with critical reasoning in huamns only, but with the use of NLMs as a possible influential variable.

This complementary perspective is essential for more precisely framing the research question, as it allows for the evaluation not only of the autonomous capabilities of NLMs but also of their role and impact in human practice, particularly within educational contexts. In doing so, the chapter explores the variables that modulate the relationship between ChatGPT usage and students' critical thinking skills, offering an integrated framework that links artificial and human performance.

The broad range of capabilities offered by next-generation language models such as ChatGPT has opened the door to numerous applications across various sectors, now reaching a user base exceeding 100 million (L. Cheng et al., 2023). In the educational domain in particular, students increasingly rely on these tools to tackle complex tasks and academic challenges (Lo, 2023; Raman et al., 2023; Zhai, 2022), making it crucial to understand how interaction with such technologies influences their CT.

4.1 Framing the Question: What We Know and Ask

4.1.1 AI and Critical Thinking in the Literature

According to Social Cognitive Theory (Bandura, 1986), behavior, cognition, and emotions are influenced by the surrounding environment, including the technology with which individuals interact (Fabio & Suriano, 2021; Zhuo et al., 2023). Several studies have highlighted how the use of technology can impact cognitive processes, including attention, memory, executive functions, and reasoning (Amez & Baert, 2020; Meltzer et al., 2023; Nakagawa et al., 2022). In particular, it has been demonstrated that technology use can affect a crucial cognitive ability in education, namely the ability to engage in complex critical thinking (X. Cheng et al., 2022; Hartanto et al., 2023; Ku et al., 2019). This process allows students to analytically evaluate information, recognize valid arguments, and develop a reflective approach to the world around them (Chang et al., 2022; Guerrero et al., 2022; Li et al., 2021).

ChatGPT can offer immediate access to a vast amount of information presented in a personalized manner, but it also raises questions about the potential influence of these interactions on students' ability to critically evaluate such information (Kasneci et al., 2023). On one hand, ChatGPT could be a valuable tool to enhance critical thinking skills, as it allows for the examination of the same topic from multiple perspectives (Rudolph et al., 2023a). This process of diversified analysis can broaden people's views, enabling them to consider the subject from different angles and enriching their thinking processes (Mollick & Mollick, 2022). However, a passive use of ChatGPT could supplant thinking processes by generating a kind of cognitive dependence. Individuals using ChatGPT might be inclined not to actively engage in cognitive processing since they can easily rely on this system that incorporates a wide range of functionalities (Cotton et al., 2023). In this perspective, excessive use of ChatGPT could undermine critical thinking ability by inducing cognitive economy. As proposed by the dual-process theory, there are two distinct reasoning systems (Stanovich & West, 2000a; Tversky & Kahneman, 1974). System 1, characterized by automatic and intuitive mental processes, occurs without the conscious control of the individual who forms quick judgments and decisions based on previous knowledge and beliefs. System 2, characterized by controlled processes, occurs under the deliberate control of the individual (De Neys & Pennycook, 2019). The latter system is slower than the former and requires more cognitive resources but allows for the logical processing of information, adhering to the standards of CT (Sowden et al., 2019b). The interaction with ChatGPT can influence these processes. The passive use of such resources could amplify heuristic thinking, compromising the ability to reason critically. Only when heuristic responses are blocked and replaced by CT can we accurately choose the information we use to form opinions or make decision (Fabio & Suriano, 2023). In consideration of the cognitive load theory (Sweller & Chandler, 1994), the use of ChatGPT presents two possibilities. On one hand, it could replace System 2 thinking processes and induce cognitive dependency, leading to cognitive overload and deterioration of critical thinking skills. Interacting with ChatGPT might lead users to avoid active engagement in complex cognitive processes, compromising their critical abilities. On the other hand, ChatGPT can alleviate cognitive load by offloading complex tasks, enhancing users' working memory efficiency and CT. By providing quick access to information and effectively organizing data, ChatGPT helps prevent cognitive overload and facilitates the creation of long-term memory schemas, thereby reducing the burden on working memory. Current literature on generative AI such as ChatGPT and its impact on critical thinking offers an intriguing and evolving perspective, maintaining a predominantly theoretical and commentary-based foundation until recent times (Chan et al., 2023; Michel-Villarreal et al., 2023; Plebani, 2023; Van den Berg & Du Plessis, 2023). Some current empirical studies appear to support the hypothesis that ChatGPT can enhance users' CT. Essel et al. (2024) conducted a significant study involving university students in Ghana, revealing that interaction with ChatGPT positively impacted students' critical, reflective, and creative thinking abilities. This suggests that the use of advanced language models like ChatGPT could represent a promising innovation for developing advanced cognitive skills among students. Guo and Lee (2023), in a specific educational context within the field of chemistry, observed significant improvements in students' ability to formulate insightful questions and analyze complex information through interaction with ChatGPT. This highlights the potential of generative AI in facilitating deeper and more critical learning processes. Minh (2024) further explored how the use of ChatGPT not only increased the depth of students' arguments but also improved the structure and complexity of their argumentative presentations. Janse van Rensburg (2024)'s study examined how ChatGPT 3.5 could be used to model CT through the generation of textual responses. It was found that targeted educational involvement is crucial for maximizing the effectiveness of this process, emphasizing the importance of active pedagogical guidance in utilizing these technologies. Particularly relevant, the findings from Avello et al. (2024) indicated a significant reduction in cognitive load in the group using ChatGPT. This suggests that the use of AI tools can facilitate CT by freeing up cognitive resources for analyzing and solving complex problems. However, few studies, such as that of Krupp et al. (2024), show that many students accept inaccurate answers and use copy-and-paste without critically evaluating information, highlighting the importance of educating students on the responsible use of AI. In contrast, Kosar et al. (2024) find no significant differences in academic performance between students who use ChatGPT and those who do not. These conflicting results underscore the need for a deeper understanding of the effects of students' interaction with ChatGPT on their critical thinking ability and the potential intervening variables in this relationship.

4.1.2 Variables Shaping the ChatGPT-Critical Thinking Relationship

The factors influencing the use of ChatGPT, and critical thinking ability are diverse, including user attitudes and trust in Artificial Intelligence (AI), representing the emotional and cognitive orientation towards AI (Ajlouni et al., 2023) Additionally, variables such as the level of engagement and in-depth knowledge specific to ChatGPT play a role (Rad et al., 2024). Trust and attitude towards AI play a fundamental role in shaping individual approaches to knowledge and engagement in this field. While attitude toward AI is defined as the set of evaluations, feelings, and predispositions that an individual or a group holds towards AI, trustworthy AI, is defined as AI that is lawful, ethical, and robust (High-Level Expert Group on Artificial Intelligence (AI HLEG, 2019). Both attitude and trust in AI is influenced by predispositions and actual experiences with AI (Glikson & Woolley, 2020). When an individual has established trust and a positive attitude towards artificial intelligence, they are more likely to be open to learning and actively exploring information related to this technology (Celik & Muukkonen, 2023). These variables could influence the acquisition of knowledge about AI and active engagement in related topics. An individual who trusts in the safety and importance of AI is more inclined to invest time and effort in understanding its principles, workings, and potential societal impacts (Choudhury & Shamszare, 2023; Michel-Villarreal et al., 2023). The ability of AI to provide immediate and personalized feedback, guidance, and support can promote an active attitude, increasing user interest and engagement (Patel & Pavlick, 2022). This trend has found resonance in the educational context, where the use of chatbots has proven effective in stimulating students' enthusiasm and attention (St-Hilaire et al., 2022), with positive implications for the learning process (Abbas et al., 2022). Chatbots can offer students a welcoming environment where they can ask questions and seek assistance without fear of judgment (Ait Baha et al., 2024). Therefore, interaction with ChatGPT could promote deeper engagement, encouraging users to mobilize their cognitive energies and apply CT. Knowledge also plays a central role in promoting responsible and effective interaction with these systems. This implies the ability to understand, use, communicate, collaborate, and critically reflect on AI applications (Long et al., 2021; Ng et al., 2021). Students with good knowledge and experience in the field of AI may maximize the benefits and reduce the risks associated with this technology, avoiding uncritical influence, and making more informed and conscious decisions (Adamopoulou & Moussiades, 2020). In turn, the growth of knowledge and active engagement in AI directly contribute to the development of CT. The growth of knowledge and active engagement in AI provides fertile ground for the cross-application of critical skills in other areas (Du et al., 2023). Analytical, synthetic, and problem-solving skills acquired can be applied in various situations. In summary, knowledge, and engagement in AI, although specific, act as catalysts for the development of broader cognitive and critical skills. AI, treated as a field of study and application, provides a platform to refine the ability to think critically, thereby contributing to a more reflective and analytical approach in problem-solving and information evaluation overall.

4.2 The Study: Measuring the Impact of ChatGPT Use on Students

In light of the unprecedented use of ChatGPT by students and the lack of prior studies specifically examining its impact on CT, the research was conducted in a stepwise manner. Therefore, before the main study, an exploratory study was carried out so that the initial findings could help refine the design of the primary research. The following paragraph will briefly outline this exploratory study and its results, before moving on, in the next section, to a more detailed description of the main study.

4.2.1 The Preliminary Explorative Study

The study comprised a total of 126 university students, with a gender distribution of 50 males, 74 females and 2 LGBTQ+. The sample size was determined using established methodologies for sample estimation, based on the relevant population proportion specific to the study. Specifically, we employed a 95% confidence interval and calculated based on a population percentage of 9%. The participants' ages ranged from 18 to 32 years, with a mean age of 24.27 years (SD=5.76). These students were recruited from classes within the departments of economics, psychology, and sciences, spanning undergraduate, graduate, and PhD programs. The strategic selection of classes aimed to ensure a diverse representation among participants, considering a range of academic disciplines. The inclusion of students from different departments and at different stages of their academic journey aimed to capture a wide array of perspectives and educational backgrounds. The recruitment process within the classes was carried out through a combination of methods, including in-class announcements, informational emails sent to students, and project presentations during lectures. The researcher personally visited university classrooms during scheduled lectures, engaging directly with present students to provide a detailed explanation of the ongoing survey. During these interactions, it was specified that completing the survey required approximately one hour. Moreover, the researcher emphasized the importance of the project and invited interested students to participate, allowing them to choose based on their availability and interest. Participants were volunteers who provided informed consent before participating in the study. They were randomly selected from within the involved classes, ensuring a balanced representation of various degree and PhD programs and academic years. This approach to participant selection was implemented to mitigate the risk of bias related to specific class or program characteristics. The recruitment process was conducted during class sessions to facilitate participation. Participants were successively assigned to one of three distinct groups: Experimental Group 1 consisted of 29 individuals who engaged with ChatGPT for more than 2 h daily; Experimental Group 2 encompassed 30 participants who interacted with ChatGPT for approximately 1 h daily; and the Control Group comprised 67 participants who had never utilized ChatGPT. To ensure rigorous matching of participants across the three study groups, demographic, and socio-occupational variables, including gender, age, and academic performance, were carefully considered and balanced (Table 1). Regarding the level of knowledge/experience, analysis of variance (ANOVA) revealed significant group differences, F (2, 123)=13.79, $p < .001$, $\eta^2 = 0.26$, signifying disparate knowledge/experience levels among the three groups. Subsequent t-tests for pairwise comparisons between the experimental groups demonstrated significant distinctions, $t(58) = 12.33$, $p < .01$, $d = 0.8$. Similar significant differences were observed between experimental group 1 and the control group, $t(95) = 9.58$, $p < .01$, $d = 0.87$, as well as between experimental group 2 and the control group, $t(96) = 11.45$, $p < .01$, $d = 0.82$. Regarding the measure of caution in ChatGPT usage, the group factor exhibited no significant effects. To understand the primary purpose for which participants use ChatGPT the two experimental groups showed similar results: no one choose it for socialization, 60% and 62% use it for Learning or socialization purposes, 5% and 4% use it for work related tasks, 20% of both for Leisure and entertainment.

Table 4.1: Characteristics of the sample

Variable	Experimental Group 1 (n=29)	Experimental Group 2 (n=30)	**Control Group (n=67)**
Gender			
Female (%)	17 (59)	18 (60)	39 (58)
Male (%)	11 (38)	12 (40)	27 (41)
LGBTQ+ (%)	1 (3)	1 (1)	–
Age			
Range	18–32	18–32	18–32
Mean ± SD	24.21 ± 5.55	23.78 ± 4.81	24.83 ± 4.82
Educational Level			
High school (%)	27 (93)	28 (93)	64 (95)
Bachelor's degree (%)	1 (3)	1 (3)	2 (3)
Master's degree (%)	1 (3)	1 (3)	1 (2)
Grade Point Average			
Mean ± SD	27.11 ± 3.27	26.90 ± 2.89	26.50 ± 3.88
Knowledge/Experience			
Mean ± SD	20.07 ± 3.93	14.81 ± 4.10	8.97 ± 3.09
Caution			
Mean ± SD	3.38 ± 1.01	3.61 ± 1.09	3.39 ± 1.40

Note: The ChatGPT Knowledge/Experience values are the result of summing the responses of the five questions of Table 2, with a range that spans from 5 to a maximum of 25.

The study was conducted in accordance with the principles established in the Declaration of Helsinki. All participants voluntarily expressed their willingness to participate in this research. Prior to the start of the study, each participant provided written informed consent. Participants were welcomed in small groups in a quiet room near their classroom and subjected to a 60-minute CT assessment. All the assessment were conducted between 9:00 AM and 12:00 PM. In this study, the level of knowledge/experience with ChatGPT was assessed through the administration of a selfreport questionnaire adapted from the USEQ tool (Bernava et al., 2021; Gil-Gómez et al., 2017) (table), while the ability to think critically was evaluated using the Critical Reasoning Assessment (Anghel et al., 2021a). An examiner provided each participant with the necessary materials, consisting of three dilemmas and their respective questions. Participants were asked to carefully read the first dilemma and respond to the corresponding questions after completing the reading. This process was then repeated for the second and third dilemmas.

Statistical analysis was performed using SPSS 24.0 software (SPSS Inc.,

Table 4.2: Questionnaire on Knowledge/Experience of ChatGPT

Questions	Response
1. Indicate your level of knowledge with ChatGPT[1]	☐1 ☐2 ☐3 ☐4 ☐5
2. Indicate your level of experience with ChatGPT[2]	☐1 ☐2 ☐3 ☐4 ☐5
3. Indicate how frequently you have ever used ChatGPT[3]	☐1 ☐2 ☐3 ☐4 ☐5
4. Indicate your level of familiarity with the parameters that can be adjusted to regulate ChatGPT's responses[4]	☐1 ☐2 ☐3 ☐4 ☐5
5. Indicate your level of knowledge regarding privacy and security factors[5]	☐1 ☐2 ☐3 ☐4 ☐5
Additional Questions:	
Indicate your level of caution when using ChatGPT. (This pertains to your degree of care or concern when interacting with ChatGPT, encompassing aspects like trust, accuracy, and potential misinterpretations.)	
Indicate the primary purpose(s) for which you use ChatGPT. You can select one or more of the following options:	
1. Work-related tasks	
2. Leisure or entertainment	
3. Socializing or casual conversation	
4. Learning or educational purposes	
5. Other (please specify): ____________________	

[1] Refers to the functioning, capabilities, and limitations of ChatGPT
[2] Refers to familiarity with the practical use of ChatGPT
[3] Refers to quantifying the use of ChatGPT over time to understand if it has been occasional or continuous
[4] Refers to knowledge of technical aspects such as customization settings
[5] Refers to knowledge of ethical aspects

Chicago, IL, USA). The preliminary measurement parameters were the mean level of knowledge / experience of ChatGPT and the level of caution in using ChatGPT. To control for differences in knowledge/experience levels and caution, ANOVA tests were performed, with the variable *groups* as a factor. Regarding the assessment of CT, the measures were the mean of correct responses (CR) for each phase of CT for each group. For the analysis of differences between groups, a MANOVA design was used with one between-subjects variable and one within-subjects variable: 3 (groups: high knowledge/experience and high caution, low knowledge/experience and high caution, low knowledge/experience and low caution) X 5 (CT dimensions: Cognitive Complexity, Reasoning Style, Openness, Nature of Knowledge, Nature of Justification). For all statistical tests, a significance level of $\alpha = 0.05$ was established. Whenever significant effects were observed, the effect size of the test was reported. We used a partial eta squared (η^2) for MANOVA and Cohen's effect size d for the t test. Additionally, probability values for repeated measures were adjusted using the Greenhouse-Geisser correction for non-sphericity.

To investigate whether university students who heavily use ChatGPT show significant improvement or deterioration in their critical thinking abilities compared to those who use it sparingly or not at all, a MANOVA was conducted with a between-subjects factor (Groups) and a within-subjects factor (dimensions of CT). Fig. 4.1 show the means (Ms) and standard deviations (SD) for the total scores for each of the five dimensions of CT for the experimental groups and the control group. The analysis revealed significant effects for the groups variable, F (2, 123)=16.43, $p < .01$, $\eta^2 = .29$. This indicates that there are overall differences in critical thinking abilities among the groups. To elucidate these distinctions further, post-hoc tests were employed: with reference to the first vs. second experimental group, the paired t-test revealed significant differences, $t(58) = 2.32$, $p < .01$, $d = 0.82$, indicating that the first experimental group showed higher critical thinking levels than the second experimental group. With reference to the first experimental vs. control group, significant differences were also observed,$t(101) = 4.11$, $p < .01$, d=0.89, with the first experimental group exhibiting higher critical thinking levels. Finally, with reference to the second experimental vs. control group, no statistically significant differences were, $t(95) = 0.51$, $p = .23$, d=0.81. These outcomes signify that the first experimental group exhibits heightened levels of CT compared to both the second experimental group and the control group. The absence of significance for the groups X dimensions interaction, F (2, 375)=0.65, $p = .81$, $\eta^2 = .002$, suggests that the superior CT performance of the first experimental group is consistent across all dimensions of CT. To specify which dimensions of CT (e.g., cognitive complexity, reasoning style, openness, nature of knowledge, nature of justification) are particularly enhanced through interaction with ChatGPT, we conducted independent samples t-tests for each dimension across the groups. Regarding cognitive complexity, reasoning style, and openness, the t-tests revealed significant differences between Experimental Group 1 and the Control Group, specifically $t(94) = 11.45$, $p < .01$, $d = 0.87$; $t(94) = 9.27$, $p < .01$, $d = 0.88$; and $t(94) = 15.71$, $p < .01$, $d = 0.90$, respectively (Fig. 4.2). In summary, the findings indicate that heavy use of ChatGPT is associated with improved critical thinking abilities among university students, particularly in the first experimental group. These improvements are consistent across all dimensions of CT, underscoring the potential of conscientious interaction with ChatGPT as a valuable tool for enhancing users' critical thinking skills. Moreover, the findings indicate that interaction with ChatGPT significantly enhances specific dimensions of CT among participants. Specifically, compared to the Control Group, participants in Experimental Group 1 demonstrated significantly increased cognitive complexity, enhanced reasoning style, and greater openness in their thinking processes. In contrast, the dimensions of 'nature of knowledge' and 'nature of justification' did not demonstrate statistically significant differences between Experimental Group 1 and the Control Group ($t(94) = 1.32$, $p = .19$ and $t(94) = 0.81$, $p = .42$, respectively). This suggests that while cognitive complexity, reasoning style, and openness are notably enhanced through interaction with ChatGPT, the effects on dimensions related to knowledge formation and justification processes were not sufficiently detectable in our study population.

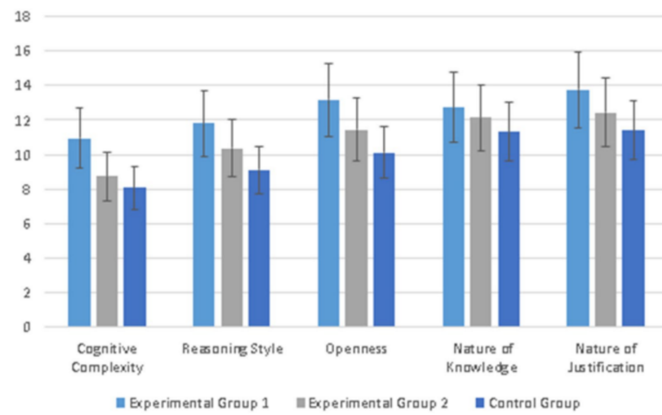


Figure 4.1: Means (Ms) and Standard Deviation (SD) of CT dimensions.

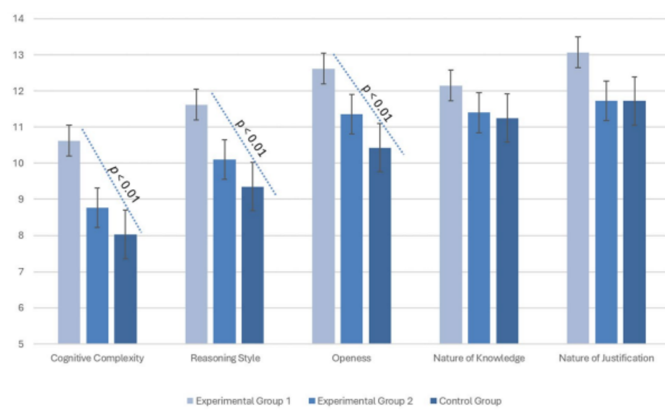


Figure 4.2: CT performance in the experimental 1, experimental 2 and control groups. The dotted trend line from experimental group 1 to the control group indicates statistical significance at $p < .01$.

4.2.2 Study Design and Hypotheses

Let us now move on to describing the main experiment, which was designed to analyze as comprehensively as possible the impact of using ChatGPT on critical thinking. Confidence in obtaining useful results from this experiment is supported by the findings of the exploratory study described in

§4.2.1, which also allowed us to refine the overall protocol employed here. This investigation is based on the analysis of various variables, including attitudes, trust, engagement, knowledge, performance in CT, and attitude. In more depth, it is hypothesized that: a) High levels of positive attitude and trust towards AI will lead to higher levels of knowledge and engagement with AI. This is supported by research from Ajlouni et al. (2023) and Celik and Muukkonen (2023), which highlight how trust and a positive attitude towards AI influence openness to learning and active exploration of AI technologies. b) An increase in knowledge and experience gained through interaction with ChatGPT will be associated with a significant improvement in CT. Studies such as those by Essel et al. (2024) and Guo and Lee (2023) demonstrate that interaction with ChatGPT can enhance critical, reflective, and creative thinking. c) Greater engagement in interactions with ChatGPT will be positively correlated with an increase in CT. Minh (2024) and Janse van Rensburg (2024) show how the use of ChatGPT can enhance the depth and structure of students' arguments, thereby improving their critical thinking abilities. d) Levels of engagement and knowledge may moderate the relationship between trust, attitude, and CT. This exploratory hypothesis suggests that a positive attitude and greater trust in AI might not only promote higher engagement and knowledge acquisition but also enhance the effect of these variables on the improvement of CT. While existing literature, such as that by Patel and Pavlick (2022) and St-Hilaire et al. (2022), indicates that trust in AI and engagement can positively influence learning, it remains unclear how and to what extent engagement and acquired knowledge interact with attitude and trust to affect CT. With these hypotheses, the aim is to provide a robust foundation for exploring the complex dynamics of interaction with ChatGPT and their possible correlations with CT in educational contexts. The underlying logic is that, while interaction with other media, such as Instagram, TikTok, and Facebook, has been shown to reduce the use of System 2 by occupying cognitive resources with superficial and impulsive activities typical of System 1, thus impairing CT (Arrivillaga et al., 2022; Dharmastuti et al., 2020; Dikarsa et al., 2020), the characteristics of AI-based chatbots, such as ChatGPT, could facilitate the activation of System 2. In line with cognitive load theory, the use of ChatGPT could reduce extraneous cognitive load, which represents cognitive resources required to manage irrelevant or poorly structured information. By reducing extraneous cognitive load, ChatGPT frees cognitive resources that can be allocated to germane cognitive load, which is directly related to learning and critical thinking activities.

4.2.3 Participants, Tasks, and Procedure

Instead of the 126 students involved in the exploratory study (see §4.2.1, the main study initially involved 241 participants; however, some of them did not complete the task due to boredom and were consequently excluded from the final analysis. The sample size was determined using established methodologies for sample estimation, based on the relevant population proportion specific to the study. Specifically, we employed a 95% confidence interval and calculated based on a university population percentage of 10% (=134). The final sample of the present study comprised 213 Italian university students, with a gender distribution of 94 males and 119 females. Participants' ages ranged from 18 to 30 years, with a mean age of 23.97 years (standard deviation = 5.02). These students were recruited from courses offered in the departments of economics, psychology, and sciences across various Italian universities, encompassing both undergraduate and graduate programs. The decision to involve students from diverse academic disciplines aimed to capture a broad range of perspectives and educational backgrounds and to ensure representativeness among participants and to avoid biases related to specific class or program characteristics. The recruitment process within the classes was implemented using a combination of approaches, including in-class announcements, informative emails sent to students, and project presentations during lectures. The researcher conducted direct visits to university classrooms during scheduled sessions, engaging personally with the attending students to deliver a detailed explanation of the ongoing survey. Students who were interested were then invited to participate, enabling them to decide based on their availability and interest.

The research was conducted in accordance with the principles outlined in the Declaration of Helsinki. Each participant voluntarily agreed to take part in this study, and the investigation commenced only after obtaining written informed consent. Participants were welcomed in small groups in a quiet room near their classroom and underwent an assessment through self-report and performance tests lasting 100 min. All assessments were conducted between 9:00 a.m. and 12:00 a.m. An examiner provided each participant with the necessary materials, consisting of three dilemmas and their respective questions. Participants were asked to carefully read the first dilemma and respond to the corresponding questions after completing the reading. This process was then repeated for the second and third dilemmas.

In this study, both self-report questionnaires and performance tests were employed. Specifically, to assess the participants' level of knowledge/experience with ChatGPT and their degree of engagement, two questionnaires were adapted from USEQ (Gil-Gómez et al., 2017). The Human-Computer Trust Measure (HCTM) developed by Gulati et al. (2019) was used to measure the level of trust, while the tool developed by Schepman and Rodway (2023) was used to measure General Attitudes towards Artificial Intelligence. Critical Thinking Attitude was measured using the Italian version of the Critical Thinking Attitude Scale (CTAS), and CT performance was investigated through the administration of the Critical Reasoning Assessment (CRA), see §2.1.4.

The *General Attitudes towards Artificial Intelligence Scale (GAAIS)* was developed by Schepman and Rodway (2023) to assess individual psychological predispositions towards Artificial Intelligence (AI). This measurement tool consisted of 20 statements, divided into two main factors: one positive and one negative, along with an attention check requiring the selection of "Strongly agree." Examples of statements from the positive scale included: "AI-based systems can contribute to making people happier" and "AI can have positive impacts on people's well-being." On the other hand, examples of statements from the negative scale were: "Organizations use AI unethically" and "I feel discomfort when thinking about future uses of AI." The scale utilized a 5-point Likert scale response mode, with the options: "Strongly Disagree," "Disagree," "Neutral," "Agree," and "Strongly Agree". To calculate the overall index of the scale, the negative items were reversed and summed with the positive ones. The internal consistency analysis of the instrument demonstrated good reliability, highlighted by a Cronbach's alpha (α) value of .89.

The *Human-Computer Trust Measure (HCTM)*is a trust scale developed by Gulati et al. (2019) focusing on assessing trust in the relationship between the user and the computer (X. Cheng et al., 2022; Pinto et al., 2022). It is a self-assessment tool consisting of 11 statements, divided into four subscales: risk perception (3 statements – reversed in this subscale), competence (3 statements), benevolence (3 statements), and reciprocity (2 statements). Individuals are asked to respond to these statements using a five-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). In this study, the goal was to assess trust in AI systems, specifically, each placeholder was replaced with the term ChatGPT. To evaluate the internal consistency analysis, the Cronbach's alpha was used ($\alpha = .92$).

To obtain an instrument for estimating the *level of knowledge/experience with ChatGPT*, a questionnaire adapted from the USEQ instrument (Bernava et al., 2021; Gil-Gómez et al., 2017) was employed. The questionnaire comprised two sub-scales: the first investigated the level of knowledge about ChatGPT, presenting three questions such as "Indicate your level of knowledge regarding privacy and security factors in GPT"; the second scale with three items referred to the level of experience with ChatGPT. An example item was "Have you used ChatGPT as support in your academic/school activities?" Participants rated each item on a 5-point Likert scale, from 1 (not at all) to 5 (very much). Internal consistency was $\alpha = .89$.

To assess the *level of engagement during the use of ChatGPT*, a questionnaire was adapted from the USEQ instrument (Bernava et al., 2021; Gil-Gómez et al., 2017) consisting of 4 items aimed at investigating the degree of active engagement and interest shown by users when interacting with the ChatGPT system. Examples of items were: "I am having a lot of fun using ChatGPT" and "I feel engaged using ChatGPT." Participants were asked to provide feedback on intensity, and the response mode was a 5-point Likert scale, ranging from 1 (not at all) to 5 (very much). From the internal consistency analysis, the Cronbach's alpha was $\alpha = .82$.

The Italian version of the Critical Reasoning Assessment (CRA) (Anghel et al., 2021b; Fabio, Plebe, Ascone, & Suriano, 2025)—see §2.1.4—was administered to objectively assess critical thinking abilities, reducing bias in responses and focusing on participants' actual performances rather than self-declarations. In the study context, responses were evaluated by two experts, achieving a high degree of agreement with a Cohen's Kappa agreement coefficient of .95. The CRA has demonstrated good reliability. For the sample of the present study, the Cronbach's alpha coefficient from the internal consistency analysis was $\alpha = .89$.

Critical Thinking Attitude Scale (CTAS; Fabio et al., 2025) was used to examine results and to assess participants' inclination to engage in a process of analytical, evaluative, and metacognitive reflection. The CTAS consists of 26 items, that measure 4 subscales "Systematicity" (nine items), "Search for Truth and Openness" (six items), "Analyticity" (four items) and "Inquisitiveness" (seven items). Participants responded on a Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), where higher scores indicated a better attitude. Example of items were: " I am able to think logically." and "I always try to delve deeper and understand things thoroughly". The CTAS demonstrated high internal consistency (Cronbach's alpha = .83) for all factors.

4.2.4 How the Data was Analyzed

To analyses the findings, we utilized SPSS 28.0 statistical software (SPSS Inc., Chicago, IL, USA). Descriptive statistics, including means and standard deviations, were provided for each variable. Pearson correlation analysis was conducted to explore the relationships among attitudes toward AI, trust, engagement, knowledge, performance, and attitude for critical thinking. Bonferroni's correction was applied to address multiple comparisons. To investigate the relationships between attitude and trust as predictive variables and engagement and knowledge as dependent variables, as well as between engagement and knowledge and critical thinking skills with critical thinking attitude, linear regression analyses were performed. Additionally, this study employed path analysis to examine the relationship among attitudes toward AI, trust, engagement, knowledge, performance, and critical thinking attitude.

4.2.5 What We Found

Table 4.3 shows the means and standard deviations of the questionnaires related to attitudes and trust toward AI, engagement, knowledge, performance, and attitude toward critical thinking. In Table 3.2, Pearson correlations between attitude and trust toward AI, engagement, knowledge, performance, and critical thinking attitude are presented. The Bonferroni correction has been applied, and the new significance level is set at (.05/15) .003. All variables show a positive and moderate correlation with performance in CT (CRA). When it comes to the attitude toward CT, only engagement and attitude towards AI exhibit correlations with it.

Table 4.3: Descriptive statistics of AI attitude, AI trust, engagement, knowledge, and critical thinking performance and attitude.

Measure	**Mean ($\pm$ SD)**
AI Attitude	62.25 ($\pm$13.96)
AI Trust	32.91 ($\pm$7.35)
Engagement	13.40 ($\pm$3.08)
Knowledge	12.99 ($\pm$5.77)
CT Performance	53.59 ($\pm$13.11)
CT Attitude	57.52 ($\pm$7.09)

To investigate the relationship between attitudes towards AI and trust in AI on levels of engagement, a linear regression was conducted, considering attitude and trust as predictor variables and engagement as the dependent variable. The data yielded significant results: respectively, $\beta = .491$, $t = 6.83$, $p < .001$, and $\beta = .282$, $t = 3.92$, $p < .001$. Together, attitude and trust explain 42.4% of the variance in predicting engagement (coefficient of determination R squared = 42.4%). The relationship between attitude

Table 4.4: Pearson's correlations between AI attitude, AI trust, engagement, knowledge and critical thinking performance and attitude.

	1	2	3	4	5	6
1. CT Performance	–					
2. CT Attitude	.415*	–				
3. AI Trust	.392*	.016	–			
4. AI Attitude	.396*	.372*	.457*	–		
5. Engagement	.401*	.394*	.494*	.605*	–	
6. Knowledge	.230*	.057	.538*	.577*	.519*	–

* $p < .001$.

and trust as predictor variables and knowledge as the dependent variable showed significant associations, with $\beta = .366$, $t = 4.94$, $p < .001$, and $\beta = .388$, $t = 5.24$, $p < .001$, respectively. Attitude and trust explain 39% of the variance in predicting knowledge (coefficient of determination R squared = 39%). Considering engagement as the independent variable and performance in critical thinking as the dependent variable, the β coefficient was .381, $t = 4.73$, $p < .001$. When considering knowledge as the independent variable and performance in critical thinking as the dependent variable, the β coefficient was .188, $t = 2.16$, $p < .05$. Together, engagement and knowledge account for 13.4% of the variance in predicting performance in critical thinking (coefficient of determination R squared = 13.4%). We conducted a path analysis to examine the relationships among AI attitude, AI trust, engagement, knowledge and critical thinking performance. The path model provided a good fit to the data ($\chi^2 = 26.32$, df = 8, $p < .001$; RMSEA = .05; CFI = .96), indicating an acceptable model fit. The path analysis illustrated in Fig. 4.3 indicates a direct relationship between attitude and trust on knowledge and engagement, and between engagement and knowledge on performance in critical thinking.

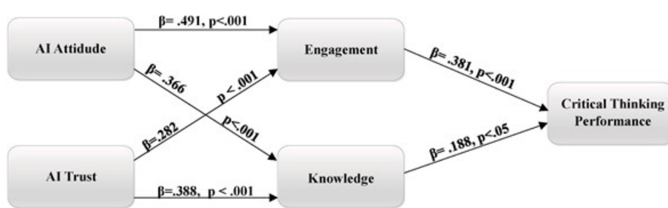


Figure 4.3: Path Analysis related to Critical Thinking Performance.

Assuming critical thinking artitudeas a dependent variable, the path model provided again a good fit to the data ($\chi^2 = 18.21$, df = 8, $p < .001$; RMSEA = .05; CFI = .96), indicating an acceptable model fit. The path analysis in Fig. 4.4 depicts a direct relationship between the same variables (trust and attitude on engagement). Knowledge, on the other hand, was found to be non-significant.

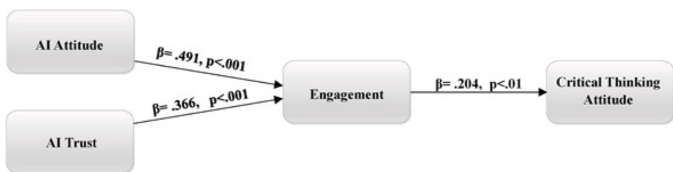


Figure 4.4: Path Analysis related to Critical Thinking Attitude.

4.3 Evidence of Positive Impact on Critical Thinking

The main objective of this study was to explore the complex dynamics of interaction with ChatGPT and their possible correlation with critical thinking in university students. Several variables were considered, such as attitudes and trust towards AI, engagement, knowledge, and experience, and their impact on both critical thinking attitude and performance.

4.3.1 Overview of Findings

The study results confirmed the first hypothesis, which posited that a positive attitude and greater trust in AI directly influence knowledge, experience, and engagement with technology. The data showed that students with a favourable attitude and high trust in ChatGPT tended to acquire more knowledge and engage more actively with the chatbot. This suggests that a positive attitude toward AI not only enhances learning but also stimulates more active participation in interacting with technology. The second and third hypotheses, which predicted a positive effect of knowledge and experience, as well as engagement on critical thinking, were confirmed. The data indicated that while the knowledge acquired was positively correlated with critical thinking performance, the effect of active engagement was more significant. This implies that although knowledge is important, active interaction with ChatGPT has a greater impact on students' critical thinking abilities. Finally, the exploratory hypothesis that levels of engagement and knowledge may moderate the relationship between trust, attitude, and critical thinking was largely confirmed. Specifically, the results demonstrated that deep engagement with the chatbot was a crucial factor for the development of critical thinking skills, suggesting the importance of the quality of interaction for thinking processes. In contrast, the acquired knowledge and experience did not have a significant impact on critical thinking attitude, indicating that the level of engagement with the technology alone is decisive in the development of critical skills.

4.3.2 Comparison with Previously Studies

The results obtained in this study offer a complex and nuanced picture compared to the existing literature on interactions with AI-based technologies such as ChatGPT. Firstly, the findings related to the influence of attitude and trust in AI on student engagement and critical performance corroborate the proposals of Ajlouni et al. (2023) and Celik and Muukkonen (2023). They suggest that a positive attitude and high trust in AI can facilitate the adoption and effective use of AI-based technologies, thereby enhancing user interaction and engagement. Our results support these observations, indicating that students with a favourable attitude towards ChatGPT and higher trust in this technology tend to be more engaged and perform better in critical thinking. Consistent with studies such as Essel et al. (2024) and Guo and Lee (2023), which found a positive impact of interaction with ChatGPT on students' critical thinking, it emerged that knowledge and experience with ChatGPT can be influential in promoting critical thinking skills. However, engagement showed a stronger impact on critical thinking performance, supporting the conclusions of recent studies like those by Minh (2024) and Janse van Rensburg (2024), which indicated that deeper engagement is crucial for improving the quality of students' critical thinking and arguments. While knowledge and experience enhance critical thinking performance by providing an informational foundation and advanced problem-solving practices, the results of this study found that these variables do not directly affect the attitude towards critical thinking, which appears to be more related to engagement. A high degree of engagement might stimulate active participation and in-depth reflection, promoting not only effectiveness in critical performance but also a positive attitude towards critical thinking itself. Therefore, while knowledge seems essential for performing critical analyses, engagement may be crucial for developing and maintaining a favourable attitude towards critical thinking (X. Cheng et al., 2022; Ku et al., 2019). The conflicting results reported by Krupp et al. (2024), who highlighted that many students accept inaccurate answers and use copy-paste without critical evaluation, suggest that while interaction with ChatGPT can facilitate critical thinking, it is insufficient without adequate engagement that stimulates deeper reflection, consideration of multiple perspectives, and the challenge of complex arguments, all essential elements for advanced critical thinking (Fabio & Suriano, 2023; Hartanto et al., 2023).

4.3.3 Theoretical Explanations

These results can be explained considering the integration among various theories. Considering Bandura (1986)'s social cognitive theory perspective, our findings suggest that positive attitude and trust toward AI can be seen as influencing factors on students' engagement and knowledge construction during interaction with ChatGPT. Considering the dual-process theory perspective (Kahneman, 2011), our results suggest that engagement with ChatGPT, along with knowledge, representing a reflective and controlled process, has proven to play a more significant role in enhancing critical thinking performance. This phenomenon can be analysed through the theory of cognitive load (Sweller & Chandler, 1994) which posits that the effectiveness of learning depends on the optimal management of available cognitive resources, dividing cognitive load into three main types: intrinsic load, extraneous load, and germane load dedicated to constructing knowledge schemas. ChatGPT appears to reduce both extraneous and intrinsic cognitive load, thereby facilitating the integration and application of tasks. This reduction allows students to allocate more cognitive resources to germane load, improving learning and understanding. Interestingly, ChatGPT could support students in processing complex information with less mental effort, thereby enhancing the overall effectiveness of the learning process. 4.4. Strengths, limitation and future direction The strengths of this study lie in its relevance during a period of widespread AI adoption. These findings contribute to the understanding of the intricate dynamics characterizing human-AI interaction in educational contexts. While AI, such as ChatGPT, can provide real-time information and support, it is the users' attitudes, trust, and active engagement that play a key role in shaping complex critical thinking skills. The results underscore the importance of considering not only technological design but also the psychological and motivational aspects of users when developing AI-based educational tools. This study takes into account the complex interaction between different variables and assesses complex critical thinking, a crucial skill in contemporary society, through both self-report measures and performance tests, thus allowing for a more in-depth and objective understanding of real performance and reported information. This study presents some significant limitations to consider in interpreting the results. Firstly, the analysis is based on a sample of university students, which, while promoting sample homogeneity, limits the generalizability of the results to other demographic groups. For instance, significant differences may emerge when examining the behavior of other age groups or different levels of education. Secondly, the cross-sectional nature of the study prevents establishing definitive causal relationships between the variables considered. Although it may be relevant for identifying correlations at a given moment, it does not allow tracing the direction or long-term effects of the interactions examined with certainty. Additionally, the assessment of interest, knowledge, and attitude to critical thinking relies on self-administered questionnaires, which could be influenced by participants' subjective perception. This introduces a potential risk of bias in the results, as responses may not fully reflect individuals' actual behavior or attitudes. To further advance understanding of these dynamics, future research could explore how variables such as education and age influence such interactions. For example, focusing on different educational levels could reveal significant variations in learning modes and the use of digital resources like ChatGPT. A longitudinal approach could be particularly fruitful in studying the evolution of interaction dynamics over time, allowing observation of changes and developments in individuals' behavior in using digital educational tools. Finally, since this study primarily focuses on students' interaction with ChatGPT in formal educational contexts, future research could expand the investigation to include informal or professional contexts. This could reveal how interaction dynamics vary based on different motivations and objectives in diverse contexts, thereby enhancing understanding of the impact of such technologies in various aspects of daily and professional life.

4.3.4 Implications and Recommendations

The current study provides a fundamental basis for further research on the interaction between humanity and Artificial Intelligence in the field of education. Understanding the dynamics involved in the use of AI-based tools, such as ChatGPT, can guide the development of more effective approaches to integrating them into education, thereby maximizing their positive impact on students' critical thinking skills. The results indicate that generative AI tools can enhance learning efficiency by facilitating access to diverse educational content tailored to individual student needs. However, it is essential to balance the benefits with the risks of this technology by establishing rigorous ethical guidelines to ensure transparency, data security, and algorithmic accountability. In the near future, it is likely that AI systems will become increasingly prominent mediators in the learning process, potentially assuming a more central role compared to traditional teachers. This perspective raises crucial considerations regarding the evolution of education and educational models. While AI systems have the potential to improve access and personalize learning, it is equally important to examine the social, emotional, and cognitive impacts of learning primarily mediated by automated technologies. Optimizing educational policies to systematically integrate AI-based tools like ChatGPT into the curriculum requires a targeted approach that actively promotes student engagement rather than merely fostering passive consumption of information. Teachers play a crucial complementary role as partners in the educational process, interpreting and guiding students' learning paths in a thorough and contextualized manner. Ongoing teacher training and structured governance of AI systems in education are essential for effectively integrating these tools into daily teaching activities. Actively involving families in the entire educational process is also crucial, for instance through targeted educational workshops that inform about the benefits and risks of using such tools. These meetings provide practical strategies to help parents actively support their children in using AI. In conclusion, while the integration of AI systems in education promises to significantly transform the learning and teaching process, it is crucial to adopt a prudent, evidence-based approach in their implementation. Only then can the full potential of emerging technologies be leveraged to improve education sustainably and equitably for all students. The use of ChatGPT can represent a valuable resource for the development of students' critical thinking skills, offering them the opportunity to explore topics from various perspectives and deepen their understanding. However, it is essential to adopt an educational approach that promotes active engagement and in-depth comprehension to avoid passive dependence on AI chatbots. The educational implications of this study suggest the need to integrate the use of AI-based chatbots into educational settings, but with a particular focus on promoting active engagement and critical understanding. Crawford et al. (2023) argued that proper guidance is necessary to safeguard the ethical use of ChatGPT in education. Teachers can play a fundamental role in guiding students in the responsible and reflective use of such resources, encouraging independent research and critical evaluation of information provided by chatbots. Additionally, students' trust in AI can be strengthened through transparency in use and training on how to critically assess received responses.

Chapter 5

Neural Language Models Beyond Education

In the preceding chapters, attention was focused on the role of NLMs in educational contexts, exploring, on the one hand, the impact of interacting with these tools on human critical thinking, and on the other, the quality of critical thinking exhibited by the models themselves in comparison to that of human users. However, to fully grasp the scope of these technologies, it is essential to move beyond the boundaries of formal learning and consider the factors that may influence the relationship between humans and language models.

This chapter aims to broaden the analysis beyond the educational dimension, by investigating the role of individual cognitive and emotional characteristics in shaping the user experience with language models. The goal is to understand which users are more likely to integrate these tools in a meaningful way into their mental, relational, and emotional lives.

Frequent interaction with chatbots exhibiting anthropomorphic characteristics can enhance users' trust and foster the development of emotional connections, thereby transforming the relationship with the conversational agent into a meaningful social bond (Araujo & Bol, 2024; Christoforakos et al., 2021; Janson, 2023; Portela & Granell-Canut, 2017). In particular, chatbots may provide socio-emotional support to individuals who experience difficulties with traditional social interactions, including those with special needs or mental health challenges (Federici et al., 2020; Meng & Dai, 2021; Pan & De Graaf, 2025; Ta et al., 2020; van Wezel et al., 2020).

Among these populations, individuals with autistic traits - defined by communication and interaction difficulties characteristic of Autism Spectrum Disorder (Diagnostic, 2013) - represent a particularly relevant group for chatbot use. Autistic traits are continuously distributed across the general population (De Groot & Van Strien, 2017; Mandy et al., 2018; Ruzich et al., 2015), and include behavioral rigidity, restricted interests, and challenges in social reciprocity (Ghanouni & Jarus, 2021; Mylett et al., 2024; X. Zhao et al., 2019). These challenges, often associated with reduced social motivation and elevated anxiety during interpersonal interactions (Bagg et al., 2024; Itskovich et al., 2021; Neuhaus et al., 2019) may lead to a preference for technology-mediated communication, which is perceived as more predictable and less threatening (F. Ali et al., 2024; Hassrick et al., 2021; Hudson et al., 2023). By offering non -judgmental, on-demand interactions, chatbots emerge as potentially valuable tools for individuals with autistic traits (Meng & Dai, 2021; Simhadri Apparao Polireddi & Kavitha, 2023; Xygkou et al., 2024).

5.1 Why Language Models Might Help Autistic Individuals

5.1.1 Theoretical Framework: the Social Motivation Theory

The Social Motivation Theory (Chevallier et al., 2012) provides a valuable theoretical framework for understanding the relationship between autistic traits and emotional attachment to chatbots. Originally developed to explain social behavior in individuals with Autism Spectrum Disorder (ASD), the theory has since been extended to account for social difficulties associated with subclinical autistic traits in the general population (Bagg et al., 2024; Foulkes et al., 2015). According to the model, individuals with ASD exhibit alterations in the brain's social reward circuitry, leading them to perceive social stimuli as less rewarding than neurotypical individuals (Blum et al., 2024; Keifer et al., 2021; Matyjek et al., 2023). In other words, social interactions are experienced as less engaging and fulfilling, resulting in reduced motivation to seek social involvement. Within this framework, interactions with chatbots may serve as a viable and accessible alternative to traditional human interaction. Given their predictability and non-threatening nature, chatbots could potentially offer a more comfortable context for engagement. It is therefore of particular interest to investigate whether, over time, such interactions may foster the development of an emotional bond.

More recent studies have questioned the validity of the Social Motivation Theory, suggesting that social withdrawal may not stem from a lack of relational desire, but rather result from other factors such as social anxiety, difficulties in understanding others, and negative interpersonal experiences (Bottini, 2018; Clements et al., 2018; Jaswal & Akhtar, 2019; Morrison et al., 2020). This line of research highlights that many autistic individuals possess a strong desire for social connection, despite experiencing real-life interactions as effortful or frustrating. In this context, chatbot interactions—characterized by their simplicity, directness, and controllability—may fulfill the need for social connection, thereby increasing the likelihood of a positive emotional response. However, existing literature has predominantly focused on the use of chatbots to enhance social skills, with relatively few studies exploring the emergence of emotional attachment to AI-based chatbots (M. R. Ali et al., 2020; Tanaka et al., 2017; Tuna, 2024).

The present study aims to address this gap by examining the role of cognitive and emotional variables that may mediate the relationship between autistic traits and emotional attachment to chatbots. As suggested by the I-PACE model (Brand et al., 2016), the interaction between individuals and technology results from the interplay of predisposing factors, affective and cognitive responses to stimuli, and the consequences of technology use. This framework supports the inclusion of cognitive and emotional variables as key mediators in understanding the development of emotional attachment to chatbots among individuals with autistic traits.

5.1.2 Cognitive Factors in Autistic Communication

Among the potential cognitive mediators of the relationship between autistic traits and emotional attachment to chatbots, particular attention has been given to Theory of Mind (ToM) and anthropomorphism. ToM—defined as the ability to attribute mental states to oneself and others, and to understand others' beliefs, intentions, and emotions (Premack & Woodruff, 1978; Tager-Flusberg & Sullivan, 2000)—is a core construct in the regulation of social interactions and in predicting the behavior of others. It plays a critical role in cognitive, social, and emotional functioning (Buehler & Weisswange, 2020; Conway & Bird, 2018; Qiu et al., 2024; Teglasi et al., 2022). In individuals with autistic traits, ToM is often impaired (Andreou & Skrimpa, 2020; Baron-Cohen et al., 1985, 1997; S. Gao et al., 2023), which limits their ability to understand others' emotions and intentions and negatively impacts interpersonal relationships (Bamicha & Drigas, 2022; Mao et al., 2023; Mazza et al., 2017; Rosello et al., 2020). This cognitive deficit is associated with increased stress and discomfort in social interactions, which require continuous inferential processing of others' mental states (Colonnesi et al., 2017; Lei & Ventola, 2018). In contrast, interactions with chatbots are typically characterized by a predictable and structurally simplified communicative format, thereby reducing the cognitive and emotional demands involved (Mygland et al., 2021; Rhim et al., 2022). Another relevant cognitive factor is anthropomorphism, defined as the tendency to attribute human-like characteristics to non-human entities (Guthrie, 1997). In the context of chatbots, anthropomorphism extends beyond the mere attribution of human traits to include the perception of cognitive capabilities such as self-control, planning, and memory, as well as affective abilities like emotional understanding and the provision of emotional support (Cao et al., 2022; Lee & Hahn, 2024). Anthropomorphism can facilitate human–machine interaction by triggering social responses in users, thereby enhancing engagement and increasing the willingness to interact with the agent (Go & Sundar, 2019; Pelau et al., 2021; Pentina et al., 2023). Specifically, the attribution of cognitive capabilities may enhance perceptions of a chatbot's competence and effectiveness, contributing to more frequent and higher-quality interactions (X. Cheng et al., 2022; R.-R. Gao et al., 2024; Jin & Youn, 2023). At the same time, affective anthropomorphism increases perceived warmth and sociability, promoting trust and facilitating the formation of emotional bonds (Q. Q. Chen & Park, 2021; Croes et al., 2023; Skjuve et al., 2021). In the absence of fulfilling social relationships, individuals may thus compensate for unmet social needs by forming affective attachments to anthropomorphized entities (Jia et al., 2024; L. Zhou et al., 2020). Several studies indicate that autistic traits are associated with a greater tendency to anthropomorphize non-human agents (Atherton & Cross, 2018; Clutterbuck et al., 2022; White & Remington, 2019). This tendency has been interpreted as a strategy to fulfill the desire for social connection within environments perceived as more controllable and predictable (Moussawi et al., 2021; Negri et al., 2019).

5.1.3 Emotional Factors in Human-AI Interaction

In addition to cognitive processes, emotional factors may serve as potential mediators in the relationship between autistic traits and emotional attachment to AI-based chatbots. In particular, social anxiety and loneliness are two emotional variables frequently associated with autistic traits, which may influence how individuals engage with conversational technologies. Individuals with autistic traits often exhibit high levels of social anxiety—a condition marked by intense discomfort in interpersonal interactions and a persistent fear of judgment or rejection (APA, 2013). This anxiety impairs the quality of social relationships and may lead to avoidant behaviors, despite an underlying desire for social connection (Alvi et al., 2022; de Lijster et al., 2018; Howell et al., 2016). Consequently, for individuals experiencing social anxiety, face-to-face interactions are frequently perceived as stressful and overwhelming, increasing the appeal of alternative, less threatening, and more controllable forms of communication—such as those enabled by AI chatbots (F. Ali et al., 2024; Dutta et al., 2018; Hutchins et al., 2021). Similarly, individuals with autistic traits are particularly vulnerable to loneliness, defined as the discrepancy between desired and actual social relationships (Reichmann, 1959). Both conditions—social anxiety and loneliness—are closely linked to the relational difficulties commonly observed in this population, with loneliness often emerging as a direct consequence of limited social inclusion (Arslan, 2023; Hunsche et al., 2022; Stark et al., 2023; Tsou et al., 2025). This condition may lead individuals to seek alternative forms of interaction capable of addressing affective deficits (De Freitas et al., 2025; Herbener & Damholdt, 2025; Jecker et al., 2024). Therefore, both social anxiety and loneliness may increase the motivation to engage with technology and chatbot interfaces, thereby facilitating the development of emotional attachment toward these tools (Hu et al., 2023; O'Day & Heimberg, 2021)

5.2 The Study: Can ChatGPT Support Communication in Autism?

5.2.1 Study Goals and Hypotheses

Due to the recent and rapid proliferation of NLM chatbots such as ChatGPT, current literature still presents significant gaps in understanding the relational dynamics that develop between users and these technologies. Unlike traditional technologies, which predominantly support passive activities such as reading or browsing, chatbots enable continuous and personalized communicative interactions (Meshram et al., 2021; Shumanov & Johnson, 2021). ChatGPT use, in particular, constitutes a direct and dynamic conversational experience, characterized by one-to-one interactions (Haqqu & Rohmah, 2024; Mukherjee et al., 2024). The present study aims to investigate the relationship between autistic traits and emotional attachment to chatbots, examining the mediating role of both cognitive and emotional factors. Specifically, four variables are considered: social anxiety, loneliness, ToM, and anthropomorphism. Based on existing literature, the following hypotheses are proposed:

H1. Higher levels of autistic traits are associated with stronger emotional attachment to AI-based chatbots.

H2. Cognitive factors—specifically, deficits in ToM and a greater tendency to anthropomorphize artificial agents—mediate the relationship between autistic traits and emotional attachment to chatbots.

H3. Emotional factors, namely social anxiety and loneliness, also mediate the relationship between autistic traits and emotional attachment to chatbots.

5.2.2 Participants, Tasks and Procedure

The sample consisted of 327 participants (139 male, 186 female, and 2 identifying as LGBTQIA+), aged between 18 and 35 years (M = 23.77; SD = 7.34). Approximately 70% of the participants were students, while the remainder were either employed or unemployed individuals. In terms of educational background, 20.18% held a Bachelor's degree, 10.09% held a Master's degree, and the remaining 69.73% had obtained either a high school diploma or a lower educational qualification. Participants were recruited through multiple channels: the questionnaire was distributed via personal contacts (e.g., WhatsApp) and social media platforms (e.g., Facebook, Instagram, and Telegram). Specifically, it was shared within online groups dedicated to ChatGPT users and Italian university students enrolled in various academic programs. This dissemination strategy enabled the collection of data from participants residing in different geographical regions of Italy: approximately 60% were from Southern Italy, while the remaining 40% were distributed across the Central and Northern regions. Table 5.1 provides a summary of the sample's demographic characteristics.

The study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki. Prior to participation, all individuals received detailed information regarding the nature and purpose of the research and provided their informed consent. The questionnaire was developed using the Google Forms web platform and subsequently distributed online. The use of social media facilitated the recruitment of a heterogeneous sample, while also reducing the time required for dissemination and data collection. Participants accessed the questionnaire through a digital link and completed it anonymously. To control for potential order effects, the sequence of the questionnaires was counterbalanced across different administrations. The tools employed in this study are as follows.

The *Autism Spectrum Quotient* (Baron-Cohen et al., 2001) is a widely used self-report questionnaire designed to assess autistic traits in adults with average or above-average intelligence (age 16+). It consists of 50 items di-

Table 5.1: Demographic information of participants (N = 327)

Variable	**Category / Value**	**N (%)**
Age	M (± SD)	23.77 (± 7.34)
Gender	Male	139 (42.51)
	Female	186 (56.88)
	LGBTQIA+	2 (0.61)
Occupation	Student	230 (70.34)
	Worker	81 (24.77)
	Unemployed	16 (4.89)
Educational level	Bachelor's degree	66 (20.18)
	Master's degree	33 (10.09)
	Diploma or other	228 (69.73)
Daily time online (h)	M (± SD)	5.79 (± 3.33)
Daily ChatGPT use (h)	M (± SD)	1.57 (± 1.05)

vided across five subdomains: social skills, attention switching, attention to detail, communication, and imagination. Example items include: "I prefer to do things with others rather than on my own," "I often become so absorbed in one thing that I lose sight of other things," "I tend to notice details that others do not," "I enjoy social chit-chat," and "When reading a story, I can easily imagine what the characters might look like." Responses are given on a 4-point Likert scale (1 = strongly agree; 4 = strongly disagree). Total scores range from 0 to 50, with higher scores indicating a greater presence of autistic traits. Statistical analyses have shown good internal consistency for the scale ($\alpha = .83$).

The *De Jong Gierveld Loneliness Scale* (De Jong Gierveld & Van Tilburg, 2006) is a shortened form of the original 11-item De Jong Gierveld Loneliness Scale (de Jong-Gierveld, 1987), developed to assess both emotional and social loneliness. The original version comprises 6 negatively worded items for emotional loneliness and 5 positively worded items for social loneliness. The short version includes 3 items from each subscale, selected to preserve the reliability and stability of the measure. Sample items include: "I experience a general sense of emptiness," "I often feel rejected," and "There are plenty of people I can rely on completely." Participants were asked to indicate how well each statement described their situation using a 3-point scale: "no" (0), "more or less" (1), and "yes" (2). The instrument has demonstrated strong internal consistency ($\alpha = .87$).

The *Social Interaction Anxiety Scale - SIAS-6* (Peters et al., 2012) is a self-report instrument designed to assess social anxiety in interactive contexts, such as one-on-one conversations, group activities, and other forms of social engagement. The scale includes 6 items, each rated on a 5-point Likert scale (0 = not at all true; 4 = extremely true). Sample items include: “I find it difficult to make eye contact with others” and “I feel tense if I am alone with another person.” The SIAS-6 has demonstrated good internal consistency (Cronbach’s $\alpha = .81$).

The *Reading the Mind in the Eyes Test - RMET* (Baron-Cohen et al., 2001), validated in Italy by Vellante et al., 2001, is a performance-based measure that assesses the affective component of ToM, namely the ability to infer others’ emotional states. The test consists of 36 photographs showing the eye region of male and female faces. For each image, participants are required to choose the emotion or mental state that best matches the expression depicted, from four possible options. The total score reflects the number of correct responses. The RMET has shown strong internal reliability (Cronbach’s $\alpha = .88$).

The *tendency to anthropomorphize ChatGPT* was assessed using a 5-item ad hoc scale developed for the purposes of this study, drawing upon the structure and content of the Attribution of Mental States Questionnaire (Miraglia et al., 2023), a widely used instrument for evaluating the attribution of mental and sensory states. Sample items adapted to the chatbot context include: “Is ChatGPT capable of thinking and/or forming opinions?” and “Is ChatGPT capable of experiencing emotions?” Items were rated on a 7-point Likert scale (1 = strongly disagree; 7 = strongly agree). The scale demonstrated excellent internal consistency, with Cronbach’s $\alpha =$ .92.

Emotional attachment to the chatbot was measured using 12 items adapted from the *Emotional Attachment Scale - EAS* (Thomson et al., 2005), originally developed to assess emotional bonding and brand loyalty. The three subscales—affection, connection, and passion—were adapted to the context by replacing references to the ”brand/product” with ”ChatGPT.” For example, the original item “I feel affection for this brand/product” was modified to “I feel affection for ChatGPT.” Items were rated on a 7-point Likert scale (1 = strongly disagree; 7 = strongly agree). The adapted scale demonstrated good internal consistency, with Cronbach’s $\alpha = .89$.

5.2.3 Statistical Data Analyses

Data analyses were conducted using SPSS version 28.0 (SPSS Inc., Chicago, IL, USA). Structural Equation Modeling (SEM) was performed using IBM SPSS AMOS to test a multiple mediation model. Autistic traits were entered as the independent variable, emotional attachment as the dependent variable, and social anxiety and loneliness (emotional factors), as well as anthropomorphism and ToM (cognitive factors), were included as parallel mediators. Model fit was assessed using several indices: chi-square (χ^2), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA) with a 90% confidence interval, and Standardized Root Mean Square Residual (SRMR). Indirect effects were tested using the bootstrap method with 5000 resamples, applying bias-corrected 95% confidence intervals.

5.2.4 Empirical Results

Descriptive statistics (means and standard deviations) and correlations among the main variables are presented in Table 5.2

Table 5.2: *Descriptive statistics and correlations between study variables*

Variable	**M (± SD)**	**1**	**2**	**3**	**4**	**5**
1. Autism	20.69 (± 6.09)	–				
2. ToM	21.99 (± 4.84)	-.12*	–			
3. Social Anxiety	12.17 (± 5.05)	.55**	.11	–		
4. Anthropomorphism	9.86 (± 10.80)	.21**	-.09	.25**	–	
5. Loneliness	9.82 (± 2.89)	.41**	.09	.56**	.23**	–
6. Emotional Attachment	22.50 (± 12.40)	.14*	-.15*	.04	.24**	.04

Note. * $p < .05$; ** $p < .001$

Autistic traits were positively correlated with social anxiety ($r = .552$, $p < .001$), loneliness ($r = .412$, $p < .001$), anthropomorphism ($r = .205$, $p < .001$), and emotional attachment (r = .142, $p < .05$), while showing a negative correlation with ToM ($r = -.117$, $p < .05$). Emotional attachment was positively associated with anthropomorphism ($r = .237$, $p < .001$), and negatively associated with ToM ($r = -.151$, $p < .05$). Significant associations were also observed between loneliness and social anxiety ($r = .563$, $p < .001$), loneliness and anthropomorphism ($r = .226$, $p < .001$), and social anxiety and anthropomorphism ($r = .254$, $p < .001$).

Mediation Model Analysis (SEM) To examine the mediating role of emotional (social anxiety and loneliness) and cognitive (anthropomorphism and ToM) factors in the relationship between autistic traits and emotional attachment, a structural equation model (SEM) with four parallel mediators was estimated. The model demonstrated a good fit to the data: $\chi^2(4) = 6.87$, $p = .142$, CFI = .991, TLI = .974, RMSEA = .039 (90% CI = [.000, .092]), SRMR = .024, indicating an adequate and well-specified structure. The results showed that autistic traits significantly predicted social anxiety ($\beta = .56$, $p < .001$), loneliness ($\beta = .41$, $p < .001$), anthropomorphism ($\beta = .21$, $p < .001$), and were negatively associated with ToM ($\beta = -.12$, $p < .05$). Social anxiety ($\beta = .08$, $p = .09$) and loneliness ($\beta = .04$, $p = .12$) did not significantly predict emotional attachment, suggesting a weak mediating role for emotional factors. In contrast, anthropomorphism significantly and positively predicted emotional attachment ($\beta = .23$, $p < .001$), while ToM exhibited a significant negative effect ($\beta = -.15$, $p < .05$), supporting the mediating role of cognitive factors. The bootstrap analysis (5000 samples) revealed that the direct effect of autistic traits on emotional attachment, controlling for the mediators, was weak but positive ($\beta = .11$, $p = .07$), suggesting partial mediation. Significant indirect effects emerged through anthropomorphism ($\beta = .048$, 95% CI [.025, .082]) and ToM ($\beta = -.018$, 95% CI [–.038, –.002]), whereas indirect effects through social anxiety ($\beta = .045$, 95% CI [–.010, .100]) and loneliness ($\beta = .016$, 95% CI [–.005, .038]) were non-significant (see Figure 1).

These results suggest that the relationship between autistic traits and emotional attachment is primarily mediated by cognitive factors, particularly the tendency to anthropomorphize and reduced ToM capacity. In contrast, emotional factors, although associated with autistic traits, do not appear to significantly mediate the link between autistic traits and emotional attachment to chatbots.

5.3 Toward Inclusive AI: Interpreting the Findings

5.3.1 Integrative Discussion

The aim of this study was to investigate how autistic traits are associated with emotional attachment to chatbots, exploring the mediating role of emotional (social anxiety and loneliness) and cognitive factors (anthropo-

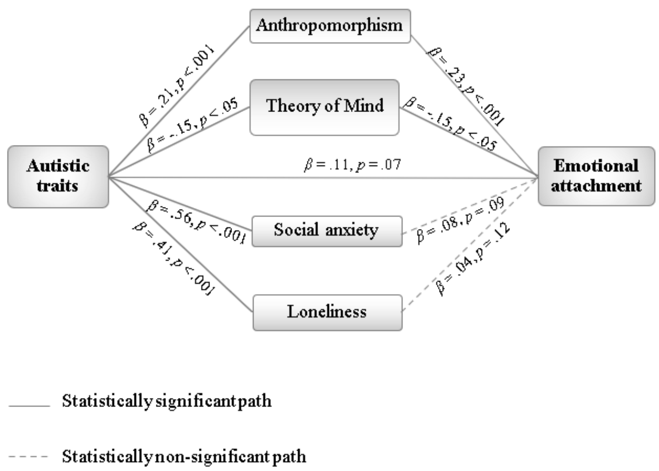


Figure 5.1: Path diagram

morphism and ToM). The results partially support the proposed hypotheses, highlighting a central role of cognitive processes in the formation of emotional bonds with artificial agents. The first hypothesis posited that higher levels of autistic traits would be associated with stronger emotional attachment to AI-based chatbots. As expected, autistic traits were positively associated with greater emotional attachment to chatbots, although the observed direct effect was marginal. This tendency can be interpreted as a preference for predictable and low-emotional-load social interactions, as often described in the literature on individuals with autism spectrum traits (Hassrick et al., 2021; Hudson et al., 2023; Tuna, 2024). However, what clearly emerges from the model is the role of cognitive mediators. The second hypothesis predicted that cognitive factors, such as deficits in ToM and a greater tendency to anthropomorphize artificial agents, would mediate the relationship between autistic traits and emotional attachment to chatbots. The findings confirmed this hypothesis, supporting the idea that attributing human-like characteristics to artificial agents and difficulty understanding others' mental states are key elements in the predisposition to develop emotional attachment to AI-based chatbots. Finally, contrary to what was predicted by the third hypothesis, emotional factors (social anxiety and loneliness), while significantly associated with autistic traits, did not directly predict emotional attachment to chatbots. This finding suggests an important distinction between the need to avoid real social situations and the tendency to form bonds with artificial entities. In other words, social anxiety and loneliness may lead individuals to withdraw from interpersonal interactions and increase their use of technology as a safe relational space (Mygland et al., 2021; Rhim et al., 2022). However, technology use driven by avoidance or isolation does not necessarily imply affective involvement or emotional attachment toward the medium itself. Forming an emotional bond requires an active interpretive and relational component — such as attributing intentionality or humanness — which is more closely tied to cognitive processes. In this sense, our findings suggest that emotional attachment to chatbots is not simply a consequence of social discomfort, but primarily emerges when users perceive the chatbot as human-like (high anthropomorphism) and have difficulty interpreting others' emotional and mental states (reduced ToM).

The findings of this study offer relevant insights, particularly in light of the Social Motivation Theory (Chevallier et al., 2012) which posits that the social difficulties observed in individuals with autistic traits stem from reduced social motivation—namely, a lower interest in or sensitivity to social stimuli. This theory suggests that autistic individuals find social interactions less rewarding due to alterations in social reward circuitry (Blum et al., 2024; Keifer et al., 2021; Matyjek et al., 2023), which in turn leads to limited engagement in relational contexts. However, our data appear to challenge a reductionist interpretation of this model. Although participants with higher autistic traits reported elevated levels of social anxiety and loneliness—consistent with previous literature—these emotional factors did not directly predict emotional attachment to chatbots. This suggests that the desire for social connection is not absent, but rather modulated by emotional and cognitive barriers that make traditional human interactions effortful or aversive. This interpretation aligns with more recent theoretical revisions to the original model (Jaswal & Akhtar, 2019; Morrison et al., 2020), which argue that reduced social participation in autism does not necessarily imply a motivational deficit. Instead, it may reflect adaptive strategies, such as withdrawal from interactions perceived as stressful, ambiguous, or potentially negative. Within this framework, interaction with chatbots—being predictable, controllable, and free from complex emotional demands—may represent an alternative route to fulfill an intact need for social connection, albeit expressed in a non-traditional manner. Thus, the present findings support a more nuanced view of social motivation: not absent, but reoriented toward forms of interaction perceived as more accessible and less emotionally taxing. In this light, chatbots may serve a compensatory function, not only on an instrumental level but also on an emotional one—but only if cognitively interpreted as valid social partners, as suggested by the key role of the cognitive mediators observed in this study.

5.3.2 Limitations and Next Steps

The present study offers an original contribution to the understanding of the psychological mechanisms underlying emotional attachment that may develop between individuals with autistic traits and AI-based chatbots, such as ChatGPT. This is an emerging and still underexplored area of research, despite the increasing diffusion and accessibility of such technologies in everyday life. A distinctive feature of this work lies in the integration of affective variables (social anxiety and loneliness) and cognitive variables (ToM and anthropomorphism), which were not only considered in isolation but also analyzed in terms of their reciprocal interactions and mediating effects. Moreover, the heterogeneous composition of the sample in terms of age, educational background, occupational status, and geographical origin contributes to enhancing the external validity of the study.

Despite these strengths, several limitations should be acknowledged. First, although the sample was diverse, its relatively small size (240 participants) limits the generalizability of the findings. Additionally, the sample consisted exclusively of Italian participants; hence, the results may not be applicable to other cultural contexts. Another limitation concerns the cross-sectional nature of the study, which does not allow for the assessment of long-term outcomes of interactions between individuals with autistic traits and chatbots. The study is also subject to temporal constraints, as the validity of its findings may be affected by the rapid evolution of technology. For instance, a new design of conversational chatbots might lead to different modes of interaction and emotional attachment compared to those observed in the current study. Finally, characteristics and difficulties associated with the autism spectrum are highly heterogeneous and may manifest differently across individuals. Consequently, the observed results may not be generalizable to the entire population; the relationship between autistic traits and emotional attachment to chatbots may in fact be influenced by individual differences or by factors not included in the current model.

In light of these limitations, future research should aim to address these issues to deepen our understanding of emotional attachment between users with autistic traits and chatbots. First, it is advisable to conduct studies with larger and more culturally diverse samples, in order to improve the generalizability of the findings and to explore potential cross-cultural differences in interaction styles and emotional bonding with chatbots. Additionally, given the cross-sectional nature of the present study, longitudinal designs are needed to examine the temporal development of relationships with artificial agents and the long-term effects on social adaptation and psychological well-being among individuals with autistic traits. Another promising avenue concerns the rapid technological advancement of conversational chatbots. Future research should investigate how the introduction of new features, advanced levels of artificial empathy, and more sophisticated interactive designs may influence the formation of emotional attachment, as these factors could significantly alter relational dynamics. Finally, given the high degree of individual variability within the autism spectrum, it would be important to more thoroughly explore clinical subgroups and specific traits (e.g., the degree of impairment in ToM or in social motivation) that may moderate interactions with chatbots. Future studies could also integrate psychological, neurobiological, and environmental factors to better understand the complexity of affective and social responses to digital technologies.

In conclusion, the present study contributes to the understanding of relational dynamics between individuals with autistic traits and AI-based chatbots. The possibility that emotional attachment may develop—mediated by cognitive factors such as ToM and anthropomorphism—suggests that interactions with chatbots are not merely functional modes of communication, but can acquire meaningful emotional value for users who often experience difficulties in traditional social interactions. Indeed, chatbots may play a role in enhancing the emotional well-being and perceived social connectedness of more vulnerable individuals, particularly in the presence of communicative challenges or in contexts of social isolation. Therefore, chatbots could be integrated as supportive tools in educational, clinical, and psychosocial settings, offering an alternative and personalized form of social connection that is responsive to the specific needs of this population.

Conclusions

The research undertaken during this PhD, and recounted in this dissertation, presents an unusual coincidence in timing. It began almost simultaneously with the emergence of the very phenomenon it set out to study: the introduction of new conversational artificial intelligence systems. The atmosphere surrounding education in 2023, at the initial explosion of this phenomenon, was marked by widespread concern. The early scientific literature on the topic was dominated by pessimism and recommendations aimed at limiting the use of these tools by students and young people in general. While fully aware of the need for caution in the face of such novel and disruptive technologies, the approach adopted at the outset of the research described here was one of openness, driven by the scientific curiosity to empirically assess the positive and negative effects of chatbots in educational contexts. As a result, one of the first articles produced from this research, *Student interaction with ChatGPT can promote complex critical thinking skills*, distinctly went against the prevailing trend. It was likely this contrarian perspective that drew significant attention to it, with the article receiving as many as 90 citations within just six months of its publication. Now, looking at the current body of literature on the effects of chatbots, it is reassuring to see that the direction taken was a sound one. Recent studies highlighting significant benefits in the use of chatbots have multiplied, and the educational sector in particular is gradually moving toward acceptance and appreciation of these new AI tools.

This thesis has examined the impact, potential, and limitations of NLMs within the contemporary educational context, with particular focus on their implications for the development of critical thinking. The interest in this topic arises from the recognition that, in recent years, the introduction and increasingly widespread adoption of artificial intelligence-based tools, such as ChatGPT, have marked a turning point in the relationship between technology and knowledge. This shift has raised unprecedented and urgent questions regarding learning modalities and the very future of cognition.

Through a comprehensive analysis, it has been highlighted how this transformation is driving a profound change in the traditional educational paradigm, challenging the conventional distinctions between natural cognition and artificial cognition. Despite the initial skepticism expressed by part of the academic and educational communities, the findings from the conducted investigations clearly demonstrate that NLMs—although grounded in stochastic and probabilistic mechanisms—can achieve performances comparable to, and in some cases even surpassing, those of humans in tasks requiring structured critical reasoning. This evidence suggests that their computational nature does not preclude a significant contribution to cognitive processes; on the contrary, it invites a more flexible and updated reconsideration of the very definitions of reasoning, understanding, and intelligence.

From this perspective, the thesis contributes to shifting the focus of the debate from the strictly ontological dimension of so-called "machine psychology" toward a more operational and functional analysis of how artificial intelligence processes and generates knowledge. Attention is thus directed to the conditions, mechanisms, and underlying logics through which NLMs handle information, revealing both convergences and divergences with human cognition. What emerges is a broad and nuanced reflection on the potential functional continuity between human minds and artificial systems: both operate based on learned patterns, recombine prior knowledge, and generate context-sensitive responses. The critical difference, however, lies in their underlying architectures: embodied, situated, and influenced by sensory experience in the case of human beings; disembodied, simulative, and algorithmic in the case of artificial models. Nonetheless, adopting this perspective does not entail disregarding the implications of future developments in NLMs. If these models already prove effective as cognitive support tools—functioning as genuine scaffolds that facilitate learning and stimulate reflective and critical practices—it is plausible to hypothesize that, with continued technological advancement, they may increasingly assume an active role, potentially even replacing, at least partially, certain cognitive functions traditionally associated with human beings.

Another area of interest lies in the growing ability of these systems to support the relational and emotional dimensions of educational experience, especially for individuals with specific needs or unconventional learning trajectories. Thanks to their capacity for personalized responses and dynamic adaptation to communicative contexts, NLMs are emerging as powerful tools for fostering more inclusive and differentiated educational practices. However, it is essential to remember that, no matter how sophisticated, these systems do not possess authentic emotions: emotion, understood as an embodied response to a real need, remains the exclusive prerogative of living beings, irreproducible by artificial entities lacking a body and direct sensory experience. This scenario invites far-reaching reflections not only within the educational domain but also at epistemological, occupational, and ethical levels. It will thus be crucial to rigorously and forward-thinkingly consider which competencies will remain uniquely human, which can be effectively delegated to artificial intelligence, and how these shifts will affect the development of critical thinking, individual autonomy, and the sense of personal and collective responsibility. Within this horizon, promoting a conscious, critical, and purpose-driven use of NLMs becomes essential—an approach capable of fully leveraging the transformative potential of these tools without ever sacrificing the centrality and irreplaceability of the human dimension in educational processes.

In conclusion, NLMs cannot be simplistically reduced to mere "stochastic parrots," nor elevated to fully autonomous minds. Rather, they represent complex cognitive artifacts, capable of redefining the very boundaries of the concept of cognition. The potential functional continuity between human and artificial intelligence is not merely a theoretical bridge between two distinct domains but constitutes a strategic and critical crossroads for radically rethinking the future of education, knowledge, and intelligence.

Bibliography

Abbas, N., Whitfield, J., Atwell, E., Bowman, H., Pickard, T., & Walker, A. (2022). Online chat and chatbots to enhance mature student engagement in higher education. *International Journal of Lifelong Education*, *41*(3), 308–326.

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research*, *78*, 1102–1134.

Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. *IFIP international conference on artificial intelligence applications and innovations*, 373–383.

Ait Baha, T., El Hajji, M., Es-Saady, Y., & Fadili, H. (2024). The impact of educational chatbot on student learning experience. *Education and Information Technologies*, *29*(8), 10153–10176.

Ajlouni, A. O., Wahba, F. A.-A., & Almahaireh, A. S. (2023). Students' attitudes towards using chatgpt as a learning tool: The case of the university of jordan. *International Journal of Interactive Mobile Technologies*, *17*(18).

Alharbi, S. M., Elfeky, A. I., & Ahmed, E. S. (2022). The effect of e-collaborative learning environment on development of critical thinking and higher order thinking skills. *Journal of Positive School Psychology*, *6*(6), 6848–6854.

Ali, F., Zhang, Q., Tauni, M. Z., & Shahzad, K. (2024). Social chatbot: My friend in my distress. *International Journal of Human–Computer Interaction*, *40*(7), 1702–1712.

Ali, M. R., Razavi, S. Z., Langevin, R., Al Mamun, A., Kane, B., Rawassizadeh, R., Schubert, L. K., & Hoque, E. (2020). A virtual conversational agent for teens with autism spectrum disorder: Experimental results and design lessons. *Proceedings of the 20th ACM international conference on intelligent virtual agents*, 1–8.

Alvi, T., Kumar, D., & Tabak, B. A. (2022). Social anxiety and behavioral assessments of social cognition: A systematic review. *Journal of affective disorders*, *311*, 17–30.

Amez, S., & Baert, S. (2020). Smartphone use and academic performance: A literature review. *International journal of educational research*, *103*, 101618.

Andreou, M., & Skrimpa, V. (2020). Theory of mind deficits and neurophysiological operations in autism spectrum disorders: A review. *Brain sciences*, *10*(6), 393.

Anggraeni, D. M., Prahani, B. K., Suprapto, N., Shofiyah, N., & Jatmiko, B. (2023). Systematic review of problem based learning research in fostering critical thinking skills. *Thinking Skills and Creativity*, *49*, 101334.

Anghel, E., Braun, H. I., Friedman, A. A., & Baez-Cruz, M. (2021a). College students' critical thinking: Assessment and interpretation. *Journal of Higher Education Theory and Practice*, *21*, 10.

Anghel, E., Braun, H. I., Friedman, A. A., & Baez-Cruz, M. (2021b). College students' critical thinking: Assessment and interpretation. *Journal of Higher Education Theory & Practice*, *21*(10).

Angius, N., Perconti, P., Plebe, A., & Acciai, A. (2025). Making sense of transformer success. *Frontiers in Artificial Intelligence*, *8*, 1509338.

Araujo, T., & Bol, N. (2024). From speaking like a person to being personal: The effects of personalized, regular interactions with conversational agents. *Computers in Human Behavior: Artificial Humans*, *2*(1), 100030.

Arrivillaga, C., Rey, L., & Extremera, N. (2022). A mediated path from emotional intelligence to problematic social media use in adolescents: The serial mediation of perceived stress and depressive symptoms. *Addictive Behaviors*, *124*, 107095.

Arslan, J. (2023). The relationship between isolation and social emotional experiences in children with autism spectrum disorder.

Atherton, G., & Cross, L. (2018). Seeing more than human: Autism and anthropomorphic theory of mind. *Frontiers in psychology*, *9*, 528.

Ausat, A. M. A., Massang, B., Efendi, M., Nofirman, & Riady, Y. (2023). Can ChatGPT replace the role of the teacher in the classroom: A fundamental analysis. *Journal on Education*, *5*, 16100–16106.

Avello, R., Gajderowicz, T., & Gómez-Rodríguez, V. G. (2024). Is chatgpt helpful for graduate students in acquiring knowledge about digital storytelling and reducing their cognitive load? an experiment. *Revista de Educación a Distancia (RED)*, *24*(78).

Bagg, E., Pickard, H., Tan, M., Smith, T. J., Simonoff, E., Pickles, A., Carter Leno, V., & Bedford, R. (2024). Testing the social motivation theory of autism: The role of co-occurring anxiety. *Journal of Child Psychology and Psychiatry*, *65*(7), 899–909.

Bamicha, V., & Drigas, A. (2022). Tom & asd: The interconnection of theory of mind with the social-emotional, cognitive development of children with autism spectrum disorder. the use of icts as an alternative form of intervention in asd. *Technium Social Sciences Journal*, *33*, 42–72.

Bandura, A. (1986). *Social cognitive theory: Social foundations of thought and action*. Prentice Hall.

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or asperger syndrome. *Journal of Child psychology and Psychiatry*, *38*(7), 813–822.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*, 37–46.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "reading the mind in the eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, *42*, 241–251.

Bartholomew, B. Y. (2020). Why AI will never surpass human intelligence. *Journal of Consciousness Exploration & Research*, *11*, 281–290.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.

Bensley, D. A. (2020). Critical thinking and the rejection of unsubstantiated claims. *Critical thinking in psychology*, *2*, 68–102.

Bernava, G., Nucita, A., Iannizzotto, G., Caprì, T., & Fabio, R. A. (2021). Proteo: A framework for serious games in telerehabilitation. *Applied Sciences*, *11*(13), 5935.

Bezanilla, M. J., Fernández-Nogueira, D., Poblete, M., & Galindo-Domínguez, H. (2019). Methodologies for teaching-learning critical thinking in higher education: The teacher's view. *Thinking Skills and Creativity*, *33*, 100584.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the Natural Academy of Science USA*, *120*, e2218523120.

Bishop, J. M. (2021). Artificial intelligence is stupid and causal reasoning will not fix it. *Frontiers in Psychology*, *11*, 513474.

Blum, K., Bowirrat, A., Sunder, K., Thanos, P. K., Hanna, C., Gold, M. S., Dennen, C. A., Elman, I., Murphy, K. T., & Makale, M. T. (2024). Dopamine dysregulation in reward and autism spectrum disorder. *Brain Sciences*, *14*(7), 733.

Boden, M. (2016). *AI: Its nature and future*. Oxford University Press.

Borowiec, K., & Castle, C. (2019). Using rater cognition to improve generalizability of an assessment of scientific argumentation. *Practical Assessment, Research, and Evaluation*, *24*(1).

Bose, R., Ogg, M., Wolmetz, M., & Ratto, C. (2024). Assessing behavioral alignment of personality-driven generative agents in social dilemma games. *Advances in Neural Information Processing Systems*.

Bottini, S. (2018). Social reward processing in individuals with autism spectrum disorder: A systematic review of the social motivation hypothesis. *Research in autism spectrum disorders*, *45*, 9–26.

Bowman, N. A. (2010). Assessing learning and development among diverse college students. *New directions for institutional research*, *145*, 53–71.

Bradeško, L., Witbrock, M., Starc, J., Herga, Z., Grobelnik, M., & Mladenisć, D. (2017). Curious cat–mobile, context-aware conversational crowdsourcing knowledge acquisition. *ACM Transactions on Information Systems*, *35*, 1–46.

Brand, M., Young, K. S., Laier, C., Wölfling, K., & Potenza, M. N. (2016). Integrating psychological and neurobiological considerations regarding the development and maintenance of specific internet-use disorders: An interaction of person-affect-cognition-execution (i-pace) model. *Neuroscience & Biobehavioral Reviews*, *71*, 252–266.

Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. *International Conference of Internet Science*, 377–392.

Brants, T., Popat, A. C., Xu, P., Och, F. J., & Dean, J. (2007). Large language models in machine translation. *Conference on Empirical Methods in Natural Language Processing*, 858–867.

Braun, H. I., Shavelson, R. J., Zlatkin-Troitschanskaia, O., & Borowiec, K. (2020). Performance assessment of critical thinking: Conceptualization, design, and implementation. *Frontiers in Education*, *5*, 156.

Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction*, *20*, 485–522.

Brookfield, S. D., & Preskill, S. (2016). *The discussion book: 50 great ways to get people talking*. John Wiley & Sons.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., & et.al., P. D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 1877–1901.

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese*, *195*, 5339–5372.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy compass*, *14*, e12625.

Buehler, M. C., & Weisswange, T. H. (2020). Theory of mind based communication for human agent cooperation. *2020 IEEE international conference on human-machine systems (ICHMS)*, 1–6.

Cao, C., Hu, Y., & Xu, H. (2022). A mind in intelligent personal assistants: An empirical study of mind-based anthropomorphism, fulfilled motivations, and exploratory usage of intelligent personal assistants. *Frontiers in psychology*, *13*, 856283.

Celik, I., & Muukkonen, H. (2023). Understanding and predicting teachers' intention to use educational chatbots. *Proceedings of the 17th International Conference of the Learning Sciences-ICLS 2023, pp. 906-909*.

Chan, M. M. K., Wong, I. S. F., Yau, S. Y., & Lam, V. S. F. (2023). Critical reflection on using chatgpt in student learning: Benefits or potential risks? *Nurse educator*, *48*(6), E200–E201.

Chang, C.-Y., Panjaburee, P., Lin, H.-C., Lai, C.-L., & Hwang, G.-H. (2022). Effects of online strategies on students' learning performance, self-efficacy, self-regulation and critical thinking in university online courses. *Educational technology research and development*, *70*(1), 185–204.

Chen, Q. Q., & Park, H. J. (2021). How anthropomorphism affects trust in intelligent personal assistants. *Industrial Management & Data Systems*, *121*(12), 2722–2737.

Chen, X., Jia, B., Peng, X., Zhao, H., Yao, J., Wang, Z., & Zhu, S. (2025). Effects of chatgpt and argument map (am)-supported online argumentation on college students' critical thinking skills and perceptions. *Education and Information Technologies*, 1–36.

Cheng, L., Wu, D., & Cao, J. (2023). Determinants of preschool teachers' knowledge-sharing behavior from a thinking style perspective. *Behavioral Sciences*, *13*(3), 230.

Cheng, X., Zhang, X., Cohen, J., & Mou, J. (2022). Human vs. ai: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms. *Information Processing & Management*, *59*(3), 102940.

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E. S., & Schultz, R. T. (2012). The social motivation theory of autism. *Trends in cognitive sciences*, *16*(4), 231–239.

Choudhury, A., & Shamszare, H. (2023). Investigating the impact of user trust on the adoption and use of chatgpt: Survey analysis. *Journal of Medical Internet Research*, *25*, e47184.

Christoforakos, L., Feicht, N., Hinkofer, S., Löscher, A., Schlegl, S. F., & Diefenbach, S. (2021). Connect with me. exploring influencing factors in a human-technology relationship based on regular chatbot use. *Frontiers in digital health*, *3*, 689999.

Clements, C. C., Zoltowski, A. R., Yankowitz, L. D., Yerys, B. E., Schultz, R. T., & Herrington, J. D. (2018). Evaluation of the social motivation hypothesis of autism: A systematic review and meta-analysis. *JAMA psychiatry*, *75*(8), 797–808.

Clutterbuck, R. A., Shah, P., Leung, H. S., Callan, M. J., Gjersoe, N., & Livingston, L. A. (2022). Anthropomorphic tendencies in autism: A conceptual replication and extension of white and remington (2019) and preliminary development of a novel anthropomorphism measure. *Autism*, *26*(4), 940–950.

Cohen, A. S., Lutzke, L., Otten, C. D., & Árvai, J. (2022). I think, therefore i act: The influence of critical reasoning ability on trust and behavior during the COVID-19 pandemic. *Risk Analysis*, *42*, 1073–1085.

Colonnesi, C., Nikolić, M., de Vente, W., & Bögels, S. M. (2017). Social anxiety symptoms in young children: Investigating the interplay of theory of mind and expressions of shyness. *Journal of abnormal child psychology*, *45*(5), 997–1011.

Conway, J. R., & Bird, G. (2018). Conceptualizing degrees of theory of mind. *Proceedings of the National Academy of Sciences*, *115*(7), 1408–1410.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*, 1183.

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *DOI: 10.1080/14703297.2023.2190148*.

Crawford, J., Cowling, M., & Allen, K.-A. (2023). Leadership is needed for ethical chatgpt: Character, assessment, and learning using artificial intelligence (ai). *Journal of University Teaching and Learning Practice*, *20*(3), 1–19.

Croes, E., Antheunis, M. L., & He, L. (2023). You go first: The effects of self-disclosure reciprocity in human-chatbot interactions. *2023 11th* *international conference on affective computing and intelligent interaction workshops and demos (ACIIW)*, 1–4.

Danaher, J., & Sætra, H. S. (2023). Mechanisms of techno-moral change: A taxonomy and overview. *Ethical Theory and Moral Practice*, *26*, 763–784.

das Graças Bruno Marietto, M., Aguiar, R. V., de Oliveira Barbosa, G., Botelho, W. T., Pimentel, E., dos Santos França, R., & da Silva, V. L. (2013). Artificial intelligence markup language: A brief tutorial. *International Journal of Computer Science & Engineering Survey*, *4*, 3.

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (2015). Psychometric considerations for the next generation of performance assessment. *Washington, DC: Center for K-12 Assessment & Performance Management, Educational Testing Service*, 1–100.

de Jong-Gierveld, J. (1987). Developing and testing a model of loneliness. *Journal of personality and social psychology*, *53*(1), 119.

de Lijster, J. M., Dieleman, G. C., Utens, E. M., Dierckx, B., Wierenga, M., Verhulst, F. C., & Legerstee, J. S. (2018). Social and academic functioning in adolescents with anxiety disorders: A systematic review. *Journal of affective disorders*, *230*, 108–117.

de Winter, J. C., Driessen, T., & Dodou, D. (2024). The use of ChatGPT for personality research: Administering questionnaires using generated personas. *Personality and Individual Differences*, *228*, 112729.

De Freitas, J., Oğuz-Uğuralp, Z., Uğuralp, A. K., & Puntoni, S. (2025). Ai companions reduce loneliness. *Journal of Consumer Research*, ucaf040.

De Groot, K., & Van Strien, J. W. (2017). Evidence for a broad autism phenotype. *Advances in Neurodevelopmental Disorders*, *1*(3), 129–140.

De Jong Gierveld, J., & Van Tilburg, T. (2006). De jong gierveld loneliness scale–short version. *European Journal of Mental Health.*

De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current directions in psychological science*, *28*(5), 503–509.

Dehouche, N. (2021). Plagiarism in the age of massive generative pre-trained transformers (GPT-3). *Ethics in Science and Environmental Politics*, *21*, 17–23.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Dharmastuti, A., Wiyono, B., Hitipeuw, I., & Rahmawati, H. (2020). Adolescent critical thinking prior to social media information sharing. *International Journal of Innovation, Creativity and Change*, *13*(10), 1195–1213.

Diagnostic, A. (2013). Statistical manual of mental disorders: Dsm-5 (ed.) washington. *DC: American Psychiatric Association*.

Dikarsa, A. A., Abdulkarim, A., & Halimi, M. (2020). Critical thinking skills of students in response political news on social media instagram. *Journal of International Conference Proceedings*, *3*(1), 46–55.

Du, X., Taylor, M., Blumofe, N., Berman, E., Lewallen, N., Geter, S., Park, H. W., Klopfer, E., Abelson, H., & Breazeal, C. (2023). Widening the global access of artificial intelligence (ai) literacy curriculum through the participation of day of ai. *Society for Information Technology & Teacher Education International Conference*, 1896–1903.

Dumas, D., & Dong, Y. (2020). Focusing the relational lens on critical thinking: How can relational reasoning support critical and analytic thinking? In *Critical thinking and reasoning* (pp. 47–63). Brill.

Dumitru, D. (2021). Critical thinking, moral courage, and sustainable development goals. universities as a driving force for societal change. In *Universities, sustainability and society: Supporting the implementation of the sustainable development goals* (pp. 171–179). Springer.

Dunn, D. S., Halonen, J. S., & Smith, R. A. (Eds.). (2008). *Teaching critical thinking in psychology – a handbook of best practices*. Blackwell Publishing.

Dutta, S., Ma, J., & De Choudhury, M. (2018). Measuring the impact of anxiety on online social interactions. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1).

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., & at al., M. A. A. (2023). "so what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642.

Earley, J. (1970). An efficient context-free parsing algorithm. *Communications of the Association for Computing Machinery*, *13*, 94–102.

Elder, L., & Paul, R. (2010). Critical thinking: Competency standards essential for the cultivation of intellectual skills, part 1. *Journal of Developmental Education*, *34*(2), 38–39.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–221.

Ennis, R. H. (2002). Goals for a critical thinking curriculum and its assessment. *Developing minds*, *3*, 44–46.

Essel, H. B., Vlachopoulos, D., Essuman, A. B., & Amankwa, J. O. (2024). Chatgpt effects on cognitive skills of undergraduate students: Receiving instant responses from ai-based conversational large language models (llms). *Computers and Education: Artificial Intelligence*, *6*, 100198.

Essien, A., Bukoye, O. T., O'Dea, X., & Kremantzis, M. (2024). The influence of AI text generators on critical thinking skills in uk business schools. *Studies in Higher Education*, *49*, 865–882.

Evans, C. M. (2020). Measuring student success skills: A review of the literature on collaboration. 21st century success skills. *National Center for the Improvement of Educational Assessment*.

Evans, J. S., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*, 223–241.

Fabio, R. A., Plebe, A., Ascone, C., & Suriano, R. (2025). Psychometric properties and validation of the critical reasoning assessment. *Personality and Individual Differences*, *246*, 113344.

Fabio, R. A., Plebe, A., & Suriano, R. (2025). AI-based chatbot interactions and critical thinking skills: An exploratory study. *Current Psychology*, *44*, 8082–8095.

Fabio, R. A., & Suriano, R. (2021). The influence of media exposure on anxiety and working memory during lockdown period in italy. *International Journal of Environmental Research and Public Health*, *18*(17), 9279.

Fabio, R. A., & Suriano, R. (2023). The influence of smartphone use on tweens' capacity for complex critical thinking. *Children*, *10*, 698.

Fabio, R. A., & Suriano, R. (2025). Thinking with humility: Investigating the role of intellectual humility in critical reasoning performance. *Personality and Individual Differences*, *244*, 113251.

Fajari, L. E. W., et al. (2021). Open-ended questions to assess critical-thinking skills in indonesian elementary school. *International Journal of Instruction*, *14*(1), 615–630.

Federici, S., de Filippis, M. L., Mele, M. L., Borsci, S., Bracalenti, M., Gaudino, G., Cocco, A., Amendola, M., & Simonetti, E. (2020). Inside pandora's box: A systematic review of the assessment of the perceived quality of chatbots for people with disabilities or special needs. *Disability and rehabilitation: assistive technology*, *15*(7), 832–837.

Firat, M. (2023). What ChatGPT means for universities: Perceptions of scholars and students. *Journal of Applied Learning & Teaching*, *6*, 22.

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models and other generative models. *Philosophy and Technology*, *36*, 15.

Foulkes, L., Bird, G., Gökçen, E., McCrory, E., & Viding, E. (2015). Common and distinct impacts of autistic traits and alexithymia on social reward. *PloS one*, *10*(4), e0121018.

Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass*, *5*(10), 914–926.

Gao, R.-R., Si, S.-W., Lin, X.-X., Wang, Y.-Z., Wang, N., Wang, J.-Y., & Luo, F. (2024). Differential relationships between autistic traits and anthropomorphic tendencies in adults and early adolescents. *Frontiers in Psychology*, *15*, 1281207.

Gao, S., Wang, X., & Su, Y. (2023). Examining whether adults with autism spectrum disorder encounter multiple problems in theory of mind: A study based on meta-analysis. *Psychonomic Bulletin & Review*, *30*(5), 1740–1758.

Gerjets, P., & Hellenthal-Schorr, T. (2008). Competent information search in the World Wide Web: Development and evaluation of a web training for pupils. *Computers in Human Behavior*, *24*, 693–715.

Gerlich, M. (2025). Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, *15*(1), 6.

Gerrish, S. (2018). *How smart machines think*. MIT Press.

Ghanouni, P., & Jarus, T. (2021). Social interaction among individuals with autism spectrum disorder: Relations with repetitive and restricted behaviors and interests. In *Repetitive and restricted behaviors and interests in autism spectrum disorders: From neurobiology to behavior* (pp. 159–173). Springer.

Gil-Gómez, J.-A., Manzano-Hernández, P., Albiol-Pérez, S., Aula-Valero, C., Gil-Gómez, H., & Lozano-Quilis, J.-A. (2017). Useq: A short questionnaire for satisfaction evaluation of virtual rehabilitation systems. *Sensors*, *17*(7), 1589.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of management annals*, *14*(2), 627–660.

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in human behavior*, *97*, 304–316.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Guerrero, J. G., Ali, S. A. A., & Attallah, D. M. (2022). The acquired critical thinking skills, satisfaction, and self confidence of nursing students and staff nurses through high-fidelity simulation experience. *Clinical Simulation in Nursing*, *64*, 24–30.

Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, *38*(10), 1004–1015.

Guo, Y., & Lee, D. (2023). Leveraging chatgpt for enhancing critical thinking skills. *Journal of Chemical Education*, *100*(12), 4876–4883.

Guthrie, S. E. (1997). Anthropomorphism: A definition and a theory. In R. W. Mitchell, N. S. Thompson, & H. L. Miles (Eds.), *Anthropomorphism, anecdotes, and animals* (pp. 50–58). State University of New York Press.

Haber, J. (2020). *Critical thinking*. MIT Press.

Hagendorff, T. (2020). Ai virtues–the missing link in putting ai ethics into practice. *arXiv preprint arXiv:2011.12750*.

Hagendorff, T. (2023). Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv*, *abs/2303.13988*.

Halpern, D. F. (2006). The nature and nurture of critical thinking. In R. J. Sternberg, H. L. Roediger, & D. F. Halpern (Eds.), *Critical thinking in psychology* (pp. 1–14). Cambridge University Press.

Halpern, D. F. (2013). The halpern critical thinking assessment: A response to the reviewers. *Inquiry: critical thinking across the disciplines*, *28*(3), 28–39.

Han, C., Du, X., Zhang, C., Lian, Y., Li, X., Gao, M., & Wang, B. (2023). Dialcot meets ppo: Decomposing and exploring reasoning paths in smaller language models. *arXiv preprint arXiv:2310.05074*.

Haqqu, R., & Rohmah, S. N. (2024). Interaction process between humans and chatgpt in the context of interpersonal communication. *Jurnal Ilmiah LISKI (Lingkar Studi Komunikasi)*, *10*(1), 23–35.

Hartanto, A., Lee, K. Y., Chua, Y. J., Quek, F. Y., & Majeed, N. M. (2023). Smartphone use and daily cognitive failures: A critical examination using a daily diary approach with objective smartphone measures. *British Journal of Psychology*, *114*(1), 70–85.

Hassrick, E. M., Holmes, L. G., Sosnowy, C., Walton, J., & Carley, K. (2021). Benefits and risks: A systematic review of information and communication technology use by autistic people. *Autism in Adulthood*, *3*(1), 72–84.

Heijltjes, A., Gog, T. V., Leppink, J., & Paas, F. (2014). Improving critical thinking: Effects of dispositions and instructions on economics students' reasoning skills. *Learning and Instruction*, *29*, 31–42.

Herbener, A. B., & Damholdt, M. F. (2025). Are lonely youngsters turning to chatbots for companionship? the relationship between chatbot usage and social connectedness in danish high-school students. *International Journal of Human-Computer Studies*, *196*, 103409.

Hinton, G., & Zemel, R. S. (1994). Autoencoders, minimum description length and Helmholtz free energy. *Advances in Neural Information Processing Systems*, 3–10.

Hinton, G. E. (2015). Where do features come from? *Cognitive Science*, *38*, 1078–1101.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *28*, 504–507.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation. *psychometrika*, *87*(2), 749–772.

Howell, A. N., Zibulsky, D. A., Srivastav, A., & Weeks, J. W. (2016). Relations among social anxiety, eye contact avoidance, state anxiety, and perception of interaction performance during a live conversation. *Cognitive behaviour therapy*, *45*(2), 111–122.

Hromei, C. D., Croce1, D., Delmonte, R., & Basili, R. (2024). La non canonica l'hai studiata? exploring LLMs and sentence canonicity in italian. *Italian Conference on Computational Linguistics*.

Hu, B., Mao, Y., & Kim, K. J. (2023). How social anxiety leads to problematic use of conversational ai: The roles of loneliness, rumination, and mind perception. *Computers in Human Behavior*, *145*, 107760.

Huang, L., Wang, Z., Yao, Y., Shan, C., Wang, H., Zhu, M., Lu, Y., Sun, P., & Zhao, X. (2015). Exploring the association between parental rearing styles and medical students' critical thinking disposition in china. *BMC Medical Education*, *15*, 88.

Hudson, N. A., Linnane, J. M., & Rayner-Smith, K. (2023). Autism and social media: A systematic review of the user experience. *Advances in Autism*, *9*(3), 201–216.

Hume, D. (1739). *A treatise of human nature* [Vol 1,2]. John Noon.

Hunsche, M. C., Cervin, M., Storch, E. A., Kendall, P. C., Wood, J. J., & Kerns, C. M. (2022). Social functioning and the presentation of anxiety in children on the autism spectrum: A multimethod, multi-informant analysis. *Journal of Psychopathology and Clinical Science*, *131*(2), 198.

Hutchins, N., Allen, A., Curran, M., & Kannis-Dymand, L. (2021). Social anxiety and online social interaction. *Australian Psychologist*, *56*(2), 142–153.

Huttenlocher, P. R. (2002). *Neural plasticity – the effects of environment on the development of the cerebral cortex*. Harvard University Press.

Hwang, J., Hand, B., & French, B. F. (2023). Critical thinking skills and science achievement: A latent profile analysis. *Thinking Skills and Creativity*, *49*, 101349.

Hyytinen, H., Nissinen, K., Ursin, J., Toom, A., & Lindblom-Ylänne, S. (2015). Problematising the equivalence of the test results of performance-based critical thinking tests for undergraduate students. *Studies in Educational Evaluation*, *44*, 1–8.

Hyytinen, H., Postareff, L., & Lindblom-Ylänne, S. (2020). Challenges in exploring individual's conceptions of knowledge and knowing: Examples of research on university students. In *Development of adult thinking* (pp. 177–190). Routledge.

Hyytinen, H., & Toom, A. (2019). Developing a performance assessment task in the finnish higher education context: Conceptual and empirical insights. *British Journal of Educational Psychology*, *89*(3), 551–563.

Hyytinen, H., Ursin, J., Silvennoinen, K., Kleemola, K., & Toom, A. (2021). The dynamic relationship between response processes and self-regulation in critical thinking assessments. *Studies in Educational Evaluation*, *71*, 101090.

Itskovich, E., Zyga, O., Libove, R. A., Phillips, J. M., Garner, J. P., & Parker, K. J. (2021). Complex interplay between cognitive ability and social motivation in predicting social skill: A unique role for social motivation in children with autism. *Autism Research*, *14*(1), 86–92.

Jaech, A., Lerer, A., Low, A., Carney, A., Neitz, A., Prokofiev, A., Sokolowsky, B., Barak, B., Minaiev, B., Hao, B., Baker, B., Houghton, B., Lugaresi, C., Voss, C., Shen, C., Orsinger, C., Kappler, D., Levy, D., Li, D., ... et al. (2024). *OpenAI o1 system card* [OpenAI blog].

Janse van Rensburg, J. (2024). Artificial human thinking: Chatgpt's capacity to be a model for critical thinking when prompted with problem-based writing activities. *Discover Education*, *3*(1), 42.

Janson, A. (2023). How to leverage anthropomorphism for chatbot service interfaces: The interplay of communication style and personification. *Computers in Human Behavior*, *149*, 107954.

Jaswal, V. K., & Akhtar, N. (2019). Being versus appearing socially uninterested: Challenging assumptions about social motivation in autism. *Behavioral and Brain Sciences*, *42*, e82.

Jecker, N. S., Sparrow, R., Lederman, Z., & Ho, A. (2024). Digital humans to combat loneliness and social isolation: Ethics concerns and policy recommendations. *Hastings Center Report*, *54*(1), 7–12.

Jia, X., Zhao, X., Tao, W., & Ahn, S. (2024). From exclusion to connection: The role of anthropomorphic chatbots in shaping customer responses post-social exclusion. *Journal of Internet commerce*, *23*(3), 205–232.

Jin, S. V., & Youn, S. (2023). Social presence and imagery processing as predictors of chatbot continuance intention in human-ai-interaction. *International Journal of Human–Computer Interaction*, *39*(9), 1874–1886.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus; Giroux.

Kasami, T. (1965). *An efficient recognition and syntax algorithm for context-free languages* (Scientific Report No. 65-758). AFCRL. Bedford (MA).

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

Keifer, C. M., Day, T. C., Hauschild, K. M., & Lerner, M. D. (2021). Social and nonsocial reward anticipation in typical development and autism spectrum disorders: Current status and future directions. *Current Psychiatry Reports*, *23*(6), 32.

Kember, D., Leung, D. Y., Jones, A., Loke, A. Y., McKay, J., Sinclair, K., Tse, H., Webb, C., Yuet Wong, F. K., Wong, M., et al. (2000). Development of a questionnaire to measure the level of reflective thinking. *Assessment & evaluation in higher education*, *25*(4), 381–395.

King, P. M., & Kitchener, K. S. (1985). *Reflective judgment theory and research: Insights into the process of knowing in the college years*

[Paper presented at the Annual Meeting of the American College Personnel Association, Boston].

King, P. M., & Kitchener, K. S. (1994). *Developing reflective judgment: Understanding and promoting intellectual growth and critical thinking in adolescents and adults. jossey-bass higher and adult education series and jossey-bass social and behavioral science series.* ERIC.

Klopfenstein, L. C., Delpriori, S., Malatini, S., & Bogliolo, A. (2017). The rise of bots: A survey of conversational interfaces, patterns, and paradigms. *Conference on Designing Interactive Systems*, 555–565.

Ko, M., Park, S. H., Park, J., & Seo, M. (2024). Investigating how large language models leverage internal knowledge to perform complex reasoning. *arXiv*, *abs/2406.19502*.

Kosar, T., Ostojić, D., Liu, Y. D., & Mernik, M. (2024). Computer science education in chatgpt era: Experiences from an experiment in a programming course for novice programmers. *Mathematics*, *12*(5), 629.

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the Natural Academy of Science USA*, *121*, e2405460121.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1090–1098.

Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J. (2024). Unreflected acceptance–investigating the negative consequences of chatgpt-assisted problem solving in physics education. In *Hhai 2024: Hybrid human ai systems for the social good* (pp. 199–212). IOS Press.

Ku, K. Y., Kong, Q., Song, Y., Deng, L., Kang, Y., & Hu, A. (2019). What predicts adolescents' critical thinking about real-life news? the roles of social media news consumption and news media literacy. *Thinking Skills and Creativity*, *33*, 100570.

Kwok, L., Bravansky, M., & Griffin, L. D. (2024). Evaluating cultural adaptability of a large language model via simulation of synthetic personas. *Conference on Language Modeling.*

Lampinen, A. K., Dasgupta, I., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J., & Hill, F. (2024). Language models, like humans show content effects on reasoning tasks. *PNAS Nexus*, *3*, pgae233.

Lantian, A., Bagneux, V., Delouvée, S., & Gauvrit, N. (2021). Maybe a free thinker but not a critical one: High conspiracy belief is associated with low critical thinking ability. *Applied Cognitive Psychology*, *35*(3), 674–684.

Larrick, R. P., Morgan, J. N., & Nisbett, R. E. (1990). Teaching the use of cost-benefit reasoning in everyday life. *Psychological Science*, *1*, 362–370.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Lee, I., & Hahn, S. (2024). On the relationship between mind perception and social support of chatbots. *Frontiers in Psychology*, *15*, 1282036.

Lei, J., & Ventola, P. (2018). Characterising the relationship between theory of mind and anxiety in children with autism spectrum disorder and typically developing children. *Research in Autism Spectrum Disorders*, *49*, 1–12.

Li, Y., Li, K., Wei, W., Dong, J., Wang, C., Fu, Y., Li, J., & Peng, X. (2021). Critical thinking, emotional intelligence and conflict management styles of medical students: A cross-sectional study. *Thinking Skills and Creativity*, *40*, 100799.

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series*, *2014*(1), 1–23.

Liu, Y., & Pásztor, A. (2022). Design and validate the employer-employee-supported critical thinking disposition inventory (2ES-CTDI) for undergraduates. *Thinking Skills and Creativity*, *46*, 101169.

Lo, C. K. (2023). What is the impact of chatgpt on education? a rapid review of the literature. *Education sciences*, *13*(4), 410.

Locke, J. (1690). *An essay concerning human understanding*. Meridian Books.

Löhn, L., Kiehne, N., Ljapunov, A., & Balke, W.-T. (2024). Is machine psychology here? on requirements for using human psychological tests on large language models. *International Natural Language Generation Conference*, 230–242.

Long, D., Blunt, T., & Magerko, B. (2021). Co-designing ai literacy exhibits for informal learning spaces. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–35.

Lyman, E. (1972). *A summary of plato curriculum and research materials* (tech. rep.). Computer-based Education Research Laboratory. Urbana (IL).

Macpherson, R., & Stanovich, K. E. (2007). Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learning and Individual Differences*, *17*, 115–127.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*, 517–540.

Mandy, W., Pellicano, L., St Pourcain, B., Skuse, D., & Heron, J. (2018). The development of autistic social traits across childhood and adolescence in males and females. *Journal of Child Psychology and Psychiatry*, *59*(11), 1143–1151.

Mao, S.-Y., Chiu, H. M., Yu, Y.-T., & Chen, K.-L. (2023). The associations of theory of mind with both general and theory-of-mind-related social interaction in children with autism spectrum disorder. *Research in Autism Spectrum Disorders*, *102*, 102107.

Marin, L. M., & Halpern, D. F. (2011). Pedagogy for developing critical thinking in adolescents: Explicit instruction produces greatest gains. *Thinking Skills and Creativity*, *6*, 1–13.

Matyjek, M., Bayer, M., & Dziobek, I. (2023). Reward responsiveness in autism and autistic traits–evidence from neuronal, autonomic, and behavioural levels. *NeuroImage: Clinical*, *38*, 103442.

Mazza, M., Mariano, M., Peretti, S., Masedu, F., Pino, M. C., & Valenti, M. (2017). The role of theory of mind on social information processing in children with autism spectrum disorders: A mediation analysis. *Journal of autism and developmental disorders*, *47*(5), 1369–1379.

Meirbekov, A., Maslova, I., & Gallyamova, Z. (2022). Digital education tools for critical thinking development. *Thinking Skills and Creativity*, *44*, 101023.

Meltzer, J. A., Kates Rose, M., Le, A. Y., Spencer, K. A., Goldstein, L., Gubanova, A., Lai, A. C., Yossofzai, M., Armstrong, S. E., & Bialystok, E. (2023). Improvement in executive function for older adults through smartphone apps: A randomized clinical trial comparing language learning and brain training. *Aging, Neuropsychology, and Cognition*, *30*(2), 150–171.

Meng, J., & Dai, Y. (2021). Emotional support from ai chatbots: Should a supportive partner self-disclose or not? *Journal of Computer-Mediated Communication*, *26*(4), 207–222.

Meshram, S., Naik, N., More, T., Kharche, S., et al. (2021). Conversational ai: Chatbots. *2021 International Conference on Intelligent Technologies (CONIT)*, 1–6.

Messick, S. (1994). The matter of style: Manifestations of personality in cognition, learning, and teaching. *Educational psychologist*, *29*(3), 121–136.

Meyer, K. (2024). Moral education through the fostering of reasoning skills. *Ethical theory and moral practice*, *27*(1), 41–55.

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). Challenges and opportunities of generative ai for higher education as explained by chatgpt. *Education sciences*, *13*(9), 856.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Minh, A. N. (2024). Leveraging chatgpt for enhancing english writing skills and critical thinking in university freshmen. *Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online)*, *3*(2), 51–62.

Minsky, M. (1954). *Neural nets and the brain-model problem* [Doctoral dissertation, Princeton University].

Miraglia, L., Peretti, G., Manzi, F., Di Dio, C., Massaro, D., & Marchetti, A. (2023). Development and validation of the attribution of mental states questionnaire (ams-q): A reference tool for assessing anthropomorphism. *Frontiers in psychology*, *14*, 999921.

Mollick, E., & Mollick, L. (2022). New modes of learning enabled by AI chatbots: Three methods and assignments. *SSRN*, *4300783*.

Monteiro, S., Sherbino, J., Sibbald, M., & Norman, G. (2020). Critical thinking, biases and dual processing: The enduring myth of generalisable skills. *Medical education*, *54*(1), 66–73.

Moore, K. (2010). The three-part harmony of adult learning, critical thinking, and decision-making. *Journal of Adult Education*, *39*(1), 1–10.

Morrison, K. E., DeBrabander, K. M., Jones, D. R., Ackerman, R. A., & Sasson, N. J. (2020). Social cognition, social skill, and social motivation minimally predict social interaction outcomes for autistic and non-autistic adults. *Frontiers in Psychology*, *11*, 591100.

Moussawi, S., Koufaris, M., & Benbunan-Fich, R. (2021). How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents. *Electronic Markets*, *31*(2), 343–364.

Mukherjee, A., Mukherjee, G., Chatterjee, A., & Tudu, B. (2024). The ultimate solution in customer engagement using chatgpt. In *Leveraging chatgpt and artificial intelligence for effective customer engagement* (pp. 28–54). IGI Global Scientific Publishing.

Mygland, M. J., Schibbye, M., Pappas, I. O., & Vassilakopoulou, P. (2021). Affordances in human-chatbot interaction: A review of the literature. *Conference on e-Business, e-Services and e-Society*, 3–17.

Mylett, M. L., Boucher, T. Q., Scheerer, N. E., & Iarocci, G. (2024). Examining the relations between social competence, autistic traits, anxiety and depression in autistic and non-autistic children. *Journal of Autism and Developmental Disorders*, *54*(8), 3094–3106.

Nakagawa, N., Ito, K., Kisara, S., Ito, T., Ohara, H., & Odanaka, K. (2022). Effect of smartphone location on pharmacy students' attention and working memory. *(No Title)*.

Negri, O., White, R. C., & Remington, A. (2019). A friendly article: The qualitative investigation of anthropomorphism in autistic and nonautistic adults. *Autism in Adulthood*, *1*(4), 286–296.

Neuhaus, E., Webb, S. J., & Bernier, R. A. (2019). Linking social motivation with social skill: The role of emotion dysregulation in autism spectrum disorder. *Development and psychopathology*, *31*(3), 931–943.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing ai literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*, 100041.

O'Day, E. B., & Heimberg, R. G. (2021). Social media use, social anxiety, and loneliness: A systematic review. *Computers in Human Behavior Reports*, *3*, 100070.

Okuhara, T., Ishikawa, H., Okada, H., Ueno, H., & Kiuchi, T. (2020). Dual-process theories to counter the anti-vaccination movement. *Applied Soft Computing Journal*, *93*, 106384.

Ollinheimo, A., & Hakkarainen, K. (2023a). Critical thinking as cooperation and its relation to mental health and social welfare. *New Ideas in Psychology*, *68*, 100988.

Ollinheimo, A., & Hakkarainen, K. (2023b). Critical thinking as cooperation and its relation to mental health and social welfare. *New Ideas in Psychology*, *68*, 100988.

OpenAI. (2023). GPT-4 technical report. *arXiv*, *abs/2303.08774*.

Orhan, A. (2022). The relationship between critical thinking and academic achievement: A meta-analysis study. *Psycho-Educational Research Reviews*, *11*(1), 283–299.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic bulletin & review*, *11*(6), 988–1010.

Özgenel, M., & Çetin, M. (2018). Development of the marmara critical thinking dispositions scale: Validity and reliability analysis. *Uluslararası Avrasya Sosyal Bilimler Dergisi*.

Pan, S., & De Graaf, M. M. (2025). Developing a social support framework: Understanding the reciprocity in human-chatbot relationship. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–13.

Papp, K. K., Huang, G. C., Clabo, L. M. L., Delva, D., Fischer, M., Konopasek, L., Schwartzstein, R. M., & Gusic, M. (2014). Milestones of critical thinking: A developmental model for medicine and nursing. *Academic Medicine*, *89*, 715–720.

Patel, R., & Pavlick, E. (2022). Mapping language models to grounded conceptual spaces. *International Conference on Learning Representations*.

Paul, R. (2005). The state of critical thinking today. *New Directions for Community Colleges*, *130*, 27–38.

Pavone, A., & Plebe, A. (2021). How neurons in deep models relate with neurons in the brain. *Algorithms*, *14*, 272.

Pelau, C., Dabija, D.-C., & Ene, I. (2021). What makes an ai device human-like? the role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, *122*, 106855.

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, *140*, 107600.

Perconti, P., & Plebe, A. (2020). Deep learning and cognitive science. *Cognition*, *203*, Article 104365.

Perconti, P., & Plebe, A. (2023). Do machines really understand meaning? (again). *Journal of Artificial Intelligence and Consciousness*, *10*, 181–206.

Perconti, P., & Plebe, A. (2025). *Turing's vision – how ai is shaping the world*. CRC Press.

Pérez, J. Q., Daradoumis, T., & Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, *28*, 1549–1565.

Peters, L., Sunderland, M., Andrews, G., Rapee, R. M., & Mattick, R. P. (2012). Development of a short form social interaction anxiety (sias) and social phobia scale (sps) using nonparametric item response theory: The sias-6 and the sps-6. *Psychological assessment*, *24*(1), 66.

Pinto, A., Sousa, S., Simões, A., & Santos, J. (2022). A trust scale for human-robot interaction: Translation, adaptation, and validation of a human computer trust scale. *Human Behavior and Emerging Technologies*, *2022*(1), 6437441.

Plebani, M. (2023). Chatgpt: Angel or demond? critical thinking is still needed.

Plebe, A., & Perconti, P. (2022). *The future of the artificial mind*. CRC Press.

Plummer, J. D., Udomprasert, P., Vaishampayan, A., Sunbury, S., Cho, K., Houghton, H., Johnson, E., Wright, E., Sadler, P. M., & Goodman, A. (2022). Learning to think spatially through curricula that embed spatial training. *Journal of Research in Science Teaching*, *59*, 1134–1168.

Portela, M., & Granell-Canut, C. (2017). A new friend in our smartphone? observing interactions with chatbots in the search of emotional engagement. *Proceedings of the XVIII International Conference on Human Computer Interaction*, 1–7.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Science*, *4*, 515–526.

Qadir, J. (2022). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. *TechRxiv*, *techrxiv.21789434*.

Qiu, X., Gao, M., Zhu, H., Li, W., & Jiang, R. (2024). Theory of mind, empathy, and prosocial behavior in children and adolescent: A meta-analysis. *Current Psychology*, *43*(22), 19690–19707.

Rad, H. S., Alipour, R., & Jafarpour, A. (2024). Using artificial intelligence to foster students' writing feedback literacy, engagement, and outcome: A case of wordtune application. *Interactive Learning Environments*, *32*(9), 5020–5040.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., and0, I. M. K., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., ... Wellman, M. (2019). Machine behaviour. *Nature*, *568*, 477–486.

Raman, R., Mandal, S., Das, P., Kaur, T., Sanjanasri, J., & Nedungadi, P. (2023). University students as early adopters of ChatGPT: Innovation diffusion study. *Research Square*, *DOI: 10.21203/rs.3.rs-2734142/v1*.

Rees, T. (2022). Non-human words: On GPT-3 as a philosophical laboratory. *Daedalus*, *151*, 168–182.

Rhim, J., Kwak, M., Gong, Y., & Gweon, G. (2022). Application of humanization to survey chatbots: Change in chatbot perception, interaction experience, and survey data quality. *Computers in Human Behavior*, *126*, 107034.

Rivas, S. F., Saiz, C., & Ossa, C. (2022). Metacognitive strategies and development of critical thinking in higher education. *Frontiers in Psychology*, *13*, 913219.

Rosello, B., Berenguer, C., Baixauli, I., García, R., & Miranda, A. (2020). Theory of mind profiles in children with autism spectrum disorder: Adaptive/social skills and pragmatic competence. *Frontiers in psychology*, *11*, 567401.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organisation in the brain. *Psychological Review*, *65*, 386–408.

Rudolph, J., Tan, S., & Tan, S. (2023a). ChatGPT bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching*, *6*, 9.

Rudolph, J., Tan, S., & Tan, S. (2023b). War of the chatbots: Bard, Bing chat, ChatGPT, Ernie and beyond. the new AI rush and its impact on higher education. *Journal of Applied Learning & Teaching*, *6*, 23.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Rumelhart, D. E., & McClelland, J. L. (1986a). On learning the past tenses of English verbs. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 216–271, Vol. 2). MIT Press.

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986b). *Parallel distributed processing: Explorations in the microstructure of cognition*. MIT Press.

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., & Baron-Cohen, S. (2015). Measuring autistic traits in the general population: A systematic review of the autism-spectrum quotient (aq) in a nonclinical population sample of 6,900 typical adult males and females. *Molecular autism*, *6*(1), 2.

Sahoo, S., & Mohammed, C. A. (2018). Fostering critical thinking and collaborative learning skills among medical students through a research protocol writing activity in the curriculum. *Korean Journal of Medical Education*, *30*, 109–118.

Santos-Meneses, L. F., Pashchenko, T., & Mikhailova, A. (2023). Critical thinking in the context of adult learning through pbl and e-learning: A course framework. *Thinking Skills and Creativity*, *49*, 101358.

Schepman, A., & Rodway, P. (2023). The general attitudes towards artificial intelligence scale (gaais): Confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human–Computer Interaction*, *39*(13), 2724–2741.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Schubert, A.-L., Ferreira, M. B., Mata, A., & Riemenschneider, B. (2021). A diffusion model analysis of belief bias: Different cognitive mechanisms explain how cognitive abilities and thinking styles contribute to conflict resolution in reasoning. *Cognition*, *211*, 104629.

Seibert, S. A. (2021). Problem-based learning: A strategy to foster generation z's critical thinking and perseverance. *Teaching and Learning in Nursing*, *16*(1), 85–88.

Sejnowski, T. J. (2018). *The deep learning revolution*. MIT Press.

Seveso, A., Poterti, D., Federici, E., Mezzanzanica, M., & Mercorio, F. (2025). ITALIC: An italian culture-aware natural language benchmark. *Conference of the Nations of the Americas Chapter of the ACL*, 1469–1478.

Shavelson, R. (2010). *Measuring college learning responsibly: Accountability in a new era*. Stanford University Press.

Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., & Marino, J. P. (2019). Assessment of university students' critical thinking: Next generation performance assessment. *International Journal of Testing*, *19*(4), 337–362.

Shoufan, A. (2023). Exploring students' perceptions of ChatGPT: Thematic analysis and follow-up survey. *IEEE Access*, *11*, 38805–38818.

Shumanov, M., & Johnson, L. (2021). Making conversations with chatbots more personalized. *Computers in human behavior*, *117*, 106627.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*, 484–489.

Simhadri Apparao Polireddi, N., & Kavitha, J. (2023). Effectiveness of automated chatbots for operational support and self-service based on fuzzy classifier for asd: N. simhadri apparao polireddi, j. kavitha. *Soft Computing*, 1–8.

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies*, *149*, 102601.

Søgaard, A. (2023). Grounding the vector space of an octopus: Word meaning from raw text. *Minds and Machines*, *33*, 33–54.

Sönmez, E., Akkaş, B. N., & Memiş, E. K. (2021). Computer-aided argument mapping for improving critical thinking: Think better! discuss better! write better! *International Journal of Contemporary Educational Research*, *7*, 291–306.

Sosu, E. M. (2013). The development and psychometric validation of a critical thinking disposition scale. *Thinking skills and creativity*, *9*, 107–119.

Sowden, P. T., Pringle, A., & Gabora, L. (2019a). The shifting sands of creative thinking: Connections to dual-process theory. In K. J. Gilhooly, L. J. Ball, & L. Macchi (Eds.), *Insight and creativity in problem solving* (pp. 40–60). Routledge.

Sowden, P. T., Pringle, A., & Gabora, L. (2019b). The shifting sands of creative thinking: Connections to dual-process theory. In *Insight and creativity in problem solving* (pp. 40–60). Routledge.

Stanovich, K. E., & West, R. F. (2000a). Advancing the rationality debate. *Behavioral and brain sciences*, *23*(5), 701–717.

Stanovich, K. E., & West, R. F. (2000b). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Science*, *23*, 665–726.

Stark, C., Groves, N. B., & Kofler, M. J. (2023). Is reduced social competence a mechanism linking elevated autism spectrum symptoms with increased risk for social anxiety? *British Journal of Clinical Psychology*, *62*(1), 129–145.

Steeves, J. K., & Harris, L. R. (Eds.). (2013). *Plasticity in sensory systems*. Cambridge University Press.

St-Hilaire, F., Vu, D. D., Frau, A., Burns, N., Faraji, F., Potochny, J., Robert, S., Roussel, A., Zheng, S., Glazier, T., et al. (2022). A new era: Intelligent tutoring systems will transform online learning for millions. *arXiv preprint arXiv:2203.03724*.

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *7*, 1285–1295.

Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? a case study using GPT-3.5. *Journal of Experimental Psychology: General*, *153*, 1066–1075.

Suriano, R., Plebe, A., Acciai, A., & Fabio, R. A. (2025). Student interaction with ChatGPT can promote complex critical thinking skills. *Learning and Instruction*, *95*, 102011.

Susnjak, T. (2022). ChatGPT: The end of online exam integrity? *arXiv*, *abs/2212.09292*.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction*, *12*, 185–233.

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *Journal of medical Internet research*, *22*(3), e16235.

Tager-Flusberg, H., & Sullivan, K. (2000). A componential view of theory of mind: Evidence from williams syndrome. *Cognition*, *76*(1), 59–90.

Tanaka, H., Negoro, H., Iwasaka, H., & Nakamura, S. (2017). Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PloS one*, *12*(8), e0182151.

Teglasi, H., Caputo, M. H., & Scott, A. L. (2022). Explicit and implicit theory of mind and social competence: A social information processing framework. *New Ideas in Psychology*, *64*, 100915.

Thomson, M., MacInnis, D. J., & Whan Park, C. (2005). The ties that bind: Measuring the strength of consumers' emotional attachments to brands. *Journal of consumer psychology*, *15*(1), 77–91.

Thornhill-Miller, B., Camarda, A., Mercier, M., Burkhardt, J.-M., Morisseau, T., Bourgeois-Bougrine, S., Vinchon, F., El Hayek, S., Augereau-Landais, M., Mourey, F., et al. (2023). Creativity, critical thinking, communication, and collaboration: Assessment, certification, and promotion of 21st century skills for the future of work and education. *Journal of Intelligence*, *11*(3), 54.

Tinn, R., Cheng, H., Gu, Y., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2023). Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, *4*(4).

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., & and Brighter Agyemang, R. H. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, *10*, 15.

Tsou, Y.-T., Nasri, M., Li, B., Blijd-Hoogewys, E. M., Baratchi, M., Koutamanis, A., & Rieffe, C. (2025). Social connectedness and loneliness in school for autistic and allistic children. *Autism*, *29*(1), 87–101.

Tuna, A. (2024). Use of chatbots to support the inclusion of people with autism spectrum disorder. In *Integrating generative ai in education to achieve sustainable development goals* (pp. 434–448). IGI Global.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*, 1124–1131.

van Wezel, M. M., Croes, E. A., & Antheunis, M. L. (2020). "i'm here for you": Can social chatbots truly support their users? a literature review. *International Workshop on Chatbot Research and Design*, 96–113.

Van den Berg, G., & Du Plessis, E. (2023). Chatgpt and generative ai: Possibilities for its contribution to lesson planning, critical thinking and openness in teacher education. *Education Sciences*, *13*(10), 998.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 6000–6010.

Vellante, M., Baron-Cohen, S., Melis, N., Marrone, M., Petretto, D., Masala, C., & Preti, A. (2001). The "reading the mind in the eyes" test: A revalidation study of the Italian version. *Cognitive Neuropsychiatry*, *18*, 326–354.

Wallace, R. S. (2008). The anatomy of A.L.I.C.E. In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the turing test philosophical and methodological issues in the quest for the thinking computer* (pp. 181–210). Springer-Verlag.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6), 186345.

Warsah, I., Morganna, R., Uyun, M., Afandi, M., & Hamengkubuwono, H. (2021). The impact of collaborative learning on learners' critical thinking skills. *International Journal of Instruction*, *14*(2), 443–460.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*, 1526–1541.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models [Survey Certification]. *Transactions on Machine Learning Research*.

Weizenbaum, J. (1966). Eliza – a computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery*, *9*, 36–45.

White, R. C., & Remington, A. (2019). Object personification in autism: This paper will be very sad if you don't read it. *Autism*, *23*(4), 1042–1045.

Xu, E., Wang, W., & Wang, Q. (2023). The effectiveness of collaborative problem solving in promoting students' critical thinking: A meta-analysis based on empirical literature. *Humanities and Social Sciences Communications*, *10*(1), 1–11.

Xygkou, A., Siriaraya, P., She, W.-J., Covaci, A., & Ang, C. S. (2024). "can i be more social with a chatbot?": Social connectedness through interactions of autistic adults with a conversational virtual human. *International Journal of Human–Computer Interaction*, *40*(24), 8937–8954.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., & Hu, X. (2023). Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *arXiv*, *abs/2304.13712*.

Yusuf, A., Bello, S., Pervin, N., & Tukur, A. K. (2024). Implementing a proposed framework for enhancing critical thinking skills in synthesizing ai-generated texts. *Thinking Skills and Creativity*, *53*, 101619.

Zaphir, L. (2021). Rationality, bias, and prejudice: Developing citizens' ability to engage in inquiry. *Educational Philosophy and Theory*, *53*(11), 1161–1170.

Zarenchnak, M., & Brown, A. (1961). Current research at Georgetown university. In H. Edmundson (Ed.), *Proceedings of the national symposium on machine translation* (pp. 63–87). Prentice Hall.

Zhai, X. (2022). ChatGPT user experience: Implications for education. *SSRN*, *4312418*.

Zhao, C., & Hu, Y. (2021). Reflections on study strategy modifications using cognitive load theory and dual processing theory in the first year of medical school. *Medical Science Educator*, *31*(2), 813–818.

Zhao, X., Li, X., Song, Y., & Shi, W. (2019). Autistic traits and prosocial behaviour in the general population: Test of the mediating effects of trait empathy and state empathic concern. *Journal of Autism and Developmental Disorders*, *49*(10), 3925–3938.

Zheng, X., Lou, J., Cao, B., Wen, X., Ji, Y., Lin, H., Lu, Y., Han, X., Zhang, D., & Sun, L. (2024). Critic-CoT: Boosting the reasoning abilities of large language model via chain-of-thoughts critic. *arXiv, abs/2408.16326*.

Zhou, L., Gao, J., Li, D., & Shum, H.-Y. (2020). The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, *46*(1), 53–93.

Zhou, X., Teng, D., & Al-Samarraie, H. (2024). The mediating role of generative ai self-regulation on students' critical thinking and problem-solving. *Education Sciences*.

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity. *arXiv preprint arXiv:2301.12867*.

Zuriguel-Pérez, E., Falcó-Pegueroles, A., Roldán-Merino, J., Agustino-Rodríguez, S., Gómez-Martín, C., & Lluch-Canut, M. (2017). Development and psychometric properties of the nursing critical thinking in clinical practice questionnaire. worldviews evidence-based nurs. 14 (4), 257–264.