

# ANNALES

*Anali za istrske in mediteranske študije*  
*Annali di Studi istriani e mediterranee*  
*Annals for Istrian and Mediterranean Studies*  
*Series Historia et Sociologia, 33, 2023, 3*





# ANNALES

**Anali za istrske in mediteranske študije**  
**Annali di Studi istriani e mediterraneei**  
**Annals for Istrian and Mediterranean Studies**

**Series Historia et Sociologia, 33, 2023, 3**

ISSN 1408-5348  
e-ISSN 2591-1775

UDK 009

Letnik 33, leto 2023, številka 3

**UREDNIŠKI ODBOR/  
COMITATO DI REDAZIONE/  
BOARD OF EDITORS:**

Roderick Bailey (UK), Gorazd Bajc, Simona Bergoč, Furio Bianco (IT), Alexander Cherkasov (RUS), Lucija Čok, Lovorka Čoralić (HR), Darko Darovec, Devan Jagodic (IT), Vesna Mikolič, Luciano Monzali (IT), Aleksej Kalc, Urška Lampe, Avgust Lešnik, John Martin (USA), Robert Matijašič (HR), Darja Mihelič, Edward Muir (USA), Žiga Oman, Vojislav Pavlović (SRB), Peter Pirker (AUT), Claudio Povoło (IT), Marijan Premović (ME), Andrej Rahten, Vida Rožac Darovec, Mateja Sedmak, Lenart Škof, Polona Tratnik, Boštjan Udovič, Marta Verginella, Špela Verovšek, Tomislav Vignjevič, Paolo Wulzer (IT), Salvator Žitko

**Glavni urednik/Redattore capo/  
Editor in chief:**

Darko Darovec

**Odgovorni urednik/Redattore  
responsabile/Responsible Editor:**

Salvator Žitko

**Uredniki/Redattori/Editors:**

Urška Lampe, Boštjan Udovič, Žiga Oman, Veronika Kos

**Gostujoča urednica/Guest Editor/  
Editrice ospite:**

Mateja Matjašič Friš

**Prevajalka/Traduttrice/Translator:**

Petra Berlot (it.)

**Oblikovalec/Progetto grafico/  
Graphic design:**

Dušan Podgornik, Darko Darovec

**Tisk/Stampa/Print:**

Založništvo PADRE d.o.o.

**Založnika/Editori/Published by:**

Zgodovinsko društvo za južno Primorsko - Koper / *Società storica del Litorale - Capodistria*® / Inštitut IRRIS za raziskave, razvoj in strategije družbe, kulture in okolja / *Institute IRRIS for Research, Development and Strategies of Society, Culture and Environment* / *Istituto IRRIS di ricerca, sviluppo e strategie della società, cultura e ambiente*®

**Sedež uredništva/Sede della redazione/  
Address of Editorial Board:**

SI-6000 Koper/Capodistria, Garibaldijeva/Via Garibaldi 18  
**e-mail:** annaleszdjp@gmail.com, **internet:** https://zdjp.si

Redakcija te številke je bila zaključena 29. 09. 2023.

**Sofinancirajo/Supporto finanziario/  
Financially supported by:**

Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS)

*Annales - Series Historia et Sociologia* izhaja štirikrat letno.

Maloprodajna cena tega zvezka je 11 EUR.

**Naklada/Tiratura/Circulation:** 300 izvodov/copie/copies

Revija *Annales, Series Historia et Sociologia* je vključena v naslednje podatkovne baze / *La rivista Annales, Series Historia et Sociologia* è inserita nei seguenti data base / *Articles appearing in this journal are abstracted and indexed in:* Clarivate Analytics (USA): Arts and Humanities Citation Index (A&HCI) in/and Current Contents / Arts & Humanities; IBZ, Internationale Bibliographie der Zeitschriftenliteratur (GER); Sociological Abstracts (USA); Referativnyi Zhurnal Viniti (RUS); European Reference Index for the Humanities and Social Sciences (ERIH PLUS); Elsevier B. V.: SCOPUS (NL); Directory of Open Access Journals (DOAJ).

To delo je objavljeno pod licenco / *Quest'opera è distribuita con Licenza* / *This work is licensed under a Creative Commons BY-NC 4.0.*



Navodila avtorjem in vsi članki v barvni verziji so prosto dostopni na spletni strani: <https://zdjp.si>.  
*Le norme redazionali e tutti gli articoli nella versione a colori sono disponibili gratuitamente sul sito: https://zdjp.si/it.*  
*The submission guidelines and all articles are freely available in color via website https://zdjp.si/en/.*



## VSEBINA / INDICE GENERALE / CONTENTS

**Andrej Mečulj:** Epigrafska spomenika s homerskim napisom na fontiku in na Pretorski palači v Kopru ..... 401  
*Monumenti epigrafici con iscrizione omerica sul fontico e sul Palazzo pretorio a Capodistria*  
*Epigraphic Monuments with a Homeric Inscription on the Fontico and on the Praetorian Palace in Koper*

**Aleksandro Burra:** I possedimenti del monastero benedettino di S. Nicolò d'Oltra (1771) ..... 413  
*The Properties of the Benedictine Monastery of S. Nicolò d'Oltra (1771)*  
*Posesti benediktinskega samostana sv. Nikolaja v Valdoltri (1771)*

**Diana Košir:** Kulturna in jezikovna dediščina Hijacinta Repiča, frančiškana pri sv. Ani v Kopru ..... 471  
*Patrimonio culturale e linguistico di Hijacint Repič, francescano del monastero di Sant'Anna di Capodistria*  
*Cultural and Linguistic Heritage of Hijacint Repič, Franciscan at St. Anne's in Koper*

**David Hazemali:** Safeguarding Liberty? Repressive Measures Against Enemy Aliens and Community Resilience in WWI United States: The Slovenian-American Experience ..... 489  
*Proteggere la libertà? Misure restrittive contro gli stranieri nemici e resilienza della comunità nella prima guerra mondiale negli Stati Uniti: L'esperienza sloveno-americana*  
*Varovanje svobode? Represivni ukrepi proti sovražnim tujcem in odpornost etnične skupnosti v ZDA med prvo svetovno vojno: izkušnja slovensko-ameriške skupnosti*

**Špela Chomicki, Renato Podbersič & Darko Friš:**  
 Nemška okupacija in organizacija Štajerske domovinske zveze v Ptujem okrožju ..... 503  
*Occupazione tedesca e l'Unione patriottica della Stiria nel Distretto di Ptuj*  
*The German Occupation and the Organisation of the Styrian Homeland Association in the Ptuj District*

**Tadeja Melanšek & Darko Friš:**  
 »Nov hram učenosti«; Ustanovitev Pedagoške akademije v Mariboru. .... 515  
*Un «Nuovo tempio del sapere»: L'istituzione dell'Accademia pedagogica di Maribor*  
*«New Temple of Learning»: The Establishment of the Pedagogical Academy in Maribor*

<b>Janez Osojnik:</b> Demosova plebiscitna pobuda: analiza spominske literature in dogajanje konec oktobra in v začetku novembra 1990 ..... 527	<b>Pirkovič Jelka:</b> The Role of Media Reports in the Democratisation of Archaeological Heritage Management – The Case of Slovenia ..... 557
<i>L'iniziativa plebiscitaria del Demos: un'analisi della letteratura memoriale e degli sviluppi a fine ottobre e inizio novembre 1990</i>	<i>Il ruolo dei servizi giornalistici nella democratizzazione della gestione del patrimonio archeologico – il caso della Slovenia</i>
<i>The Demos Plebiscite Initiative: An Analysis of the Memoirs and the Developments at the End of October and Beginning of November 1990</i>	<i>Vloga medijskih poročil pri demokratizaciji upravljanja arheološke dediščine – primer Slovenije</i>
<b>Dejan Valentinčič:</b> Primerjava (ustavno)pravnega položaja, udejanjanja manjšinskih pravic in javnega financiranja skupnosti Slovencev v državah z območja nekdanje SFRJ in skupnosti konstitutivnih narodov nekdanje SFRJ v Sloveniji ..... 537	<b>Igor Ivanović:</b> Can AI-assisted Essay Assessment Support Teachers? A Cross-sectional Mixed-methods Research Conducted at the University of Montenegro ..... 571
<i>Paragone tra lo statuto costituzionale-legale e l'implementazione dei diritti delle minoranze, nonché tra il finanziamento pubblico alle comunità slovene nelle repubbliche dell'ex Jugoslavia, e alle comunità delle nazioni costituenti dell'ex Jugoslavia residenti in Slovenia</i>	<i>Può la valutazione dei saggi con l'aiuto dell'intelligenza artificiale sostenere gli insegnanti? Uno studio trasversale con l'uso di metodi misti condotto presso l'Università del Montenegro</i>
<i>A Comparison of the Constitutional-Legal Status, the Implementation of Minority Rights and the Public Financing between Slovenian Communities in the Republics of Former Yugoslavia and Communities of Constitutive Nations of Yugoslavia in Slovenia</i>	<i>Ali lahko ocenjevanje esejev s pomočjo umetne inteligence reši profesorje? Navzkrižna raziskava z uporabo mešanih metod izvedena na Univerzi v Črni gori</i>
	Kazalo k slikam na ovitku ..... 591
	<i>Indice delle foto di copertina</i> ..... 591
	<i>Index to images on the cover</i> ..... 591

received: 2022-06-22

DOI 10.19233/ASHS.2023.30

## CAN AI-ASSISTED ESSAY ASSESSMENT SUPPORT TEACHERS? A CROSS-SECTIONAL MIXED-METHODS RESEARCH CONDUCTED AT THE UNIVERSITY OF MONTENEGRO

*Igor IVANOVIĆ*

University of Montenegro, Faculty of Philology, Danila Bojovića bb, 81400 Nikšić, Montenegro  
e-mail: iggybosnia@ucg.ac.me

### ABSTRACT

*In this study, we will try to answer the question if an AI language model can provide teachers with essay assessment solutions that are on a par with the solutions provided by experienced professors. We designed a study with the aim of comparing the essay assessment outputs of the AI language model and three of our colleagues working at the University of Montenegro. The main aim of this paper is to investigate if this AI language model can be a viable teachers' assistance tool that provides immediate and meaningful feedback to teachers and students. Our hypothesis is, with some caveats, that the AI language model is more than a viable and useful tool, capable of providing meaningful and immediate feedback, greatly reducing the assessment time, and thus helping the teachers become more efficient and consistent. We will compare the results of 78 essays assessed by three teachers with the results provided by ChatGPT and see where the two sets of results converge or diverge in terms of their individual and overall scores.*

**Keywords:** ChatGPT, automated grading, AI language models, essay assessment, essay feedback, assessment metrics, natural language processing

## PUÒ LA VALUTAZIONE DEI SAGGI CON L'AIUTO DELL'INTELLIGENZA ARTIFICIALE SOSTENERE GLI INSEGNANTI? UNO STUDIO TRASVERSALE CON L'USO DI METODI MISTI CONDOTTO PRESSO L'UNIVERSITÀ DEL MONTENEGRO

### SINTESI

*Attraverso questo studio, cercheremo di rispondere alla domanda se il modello linguistico di intelligenza artificiale che abbiamo utilizzato può fornire agli insegnanti una qualità di valutazione dei saggi paragonabile a quella fornita da insegnanti esperti. Abbiamo sviluppato uno studio che confronta i risultati della valutazione dei saggi generati tramite il modello linguistico AI con i risultati della valutazione di tre nostri colleghi che lavorano presso l'Università del Montenegro. L'obiettivo principale di questo studio è determinare se questo modello linguistico di intelligenza artificiale può diventare un utile strumento di supporto, in grado di fornire un feedback immediato e rilevante sia a insegnanti che a studenti. La nostra ipotesi, con alcune riserve, è che il modello linguistico di intelligenza artificiale sia uno strumento più che utile e fruibile, in grado di fornire feedback immediati e rilevanti, riducendo significativamente i tempi di valutazione e aiutando così gli insegnanti a essere più efficienti e coerenti. Confronteremo i risultati di 78 saggi valutati da tre insegnanti con i risultati ottenuti tramite ChatGPT e osserveremo dove i due gruppi di risultati coincidono o differiscono in termini di punteggi individuali e complessivi.*

**Parole chiave:** ChatGPT, valutazione automatica, modelli linguistici di intelligenza artificiale, valutazione del saggio, feedback sul saggio, metriche di valutazione, elaborazione del linguaggio naturale

## INTRODUCTION AND RESEARCH RATIONALE

Essay assessment has been, in one form or another, present in academia for centuries. Essays have been changing their form, and so have the criteria, but it has always been the role of a teacher or an evaluator to grade those essays and provide some feedback. Based on our experience and the experience of many other colleagues, repetitive essay assessment quickly becomes a tedious and never-ending task. For instance, first-year faculty professors will easily have more than 50, sometimes closer to 100 students to grade, which may take days to complete due to their private and professional obligations. Sometimes it is the topic, which is not inspiring, sometimes it is all about students' generic answers or bad handwriting and sometimes it is the sheer repetitive nature of this form of assessment and grading. However, our story goes beyond simple boredom. It has been suggested by some research (Erturk et al., 2022) that this assessment-induced boredom or "mental fatigue" (Grillon et al., 2015) adversely impacts grading reliability and validity leading to incrementally lower grades (Marcora et al., 2009), which is of concern since some teachers spend decades throughout their careers performing essay assessments. Even some almost century-old research papers found a similar negative correlation between boredom, mental fatigue and the scores given by teachers. Dexter (1935, 665) found that "As the scoring proceeds the variability is not erratic but tends either towards greater severity or increasing leniency." These findings are the real-world result corroborating one of the tenets of the functional theory of boredom which posits that boredom arises when our mind wants but is unable to shift itself to a more engaging and desirable activity (Elpidorou, 2022). This may lead to frustration, dissatisfaction, and affective disengagement (Mizuno et al., 2011), which, in turn, may affect a teacher's mental capability to assess essays consistently and fairly. The functional theory suggests that boredom arises when there is a discrepancy between a person's desired level of stimulation and their actual level of stimulation. When individuals are under-stimulated or engaged in repetitive and monotonous tasks, they may experience feelings of boredom as a way of signalling the need for change or variety. Boredom, if left unchecked, may have multiple negative effects on teachers' performance in terms of essay assessment. Let us elaborate more on the most salient ones. Boredom can lead to a decline in attention and engagement. When teachers are bored while reading and evaluating essays, their focus and concentration may wane, making it more challenging to maintain consistent and thorough assessment standards. They may be more likely to skim through essays or become

disinterested in the content, potentially resulting in biased or superficial evaluations. Furthermore, boredom can trigger negative affective states, leading to a negative bias in essay assessment. When teachers are bored, they may be more inclined to view essays more critically or harshly, finding flaws or weaknesses more easily. This bias can influence the grading process and result in lower scores and/or unfair evaluations. If teachers find the task monotonous or uninteresting, they may approach it with a lack of enthusiasm or investment. Consequently, they might rush through assessments or become less committed to providing constructive feedback, potentially impacting the quality and usefulness of the assessment process. Lastly, boredom can limit cognitive flexibility, hindering teachers' ability to consider alternative perspectives or appreciate the nuances of student essays. When bored, individuals tend to seek immediate stimulation or distraction, which can lead to a more rigid and narrow-minded approach to assessment. This rigidity may prevent teachers from fully understanding and appreciating the unique qualities or originality in students' work. By acknowledging the impact of boredom on teachers' assessment performance, our study recognises a real-world problem that can affect grading consistency, fairness, and the overall quality of feedback provided to students.

Thus, based on the abovementioned research from Erturk et al. (2022) and the notions put forward by the functional theory of boredom, we designed a study to answer the following question: If boredom is (an inevitable?) part of essay assessment and may negatively affect scoring trustworthiness, can an AI language model help alleviate the situation and speed up the assessment process? To answer this question, we needed to find an AI language model suitable for our needs. To find a suitable AI language model, we conducted extensive internet and literary research intending to find a model which would meet at least the following criteria:

- It needed to be user-friendly for the teachers, meaning no need for coding or any other similarly complex task.
- It needed to be easily deployable, meaning the AI language model needed to be easily distributed, accessed, and used by teachers.
- It needed to be suitable for language-related tasks such as offering feedback, simulating human-like conversation, arguing its point of view, etc.

Based on these criteria and with the help of the information elicited from some other studies (Floridi & Chiriatti, 2020; Dehouche, 2021; Gao, 2021) in which the authors used the services of an AI language model, and with the support from our

colleagues from the Computer Science Department at the Faculty of Science and Mathematics of the University of Montenegro, we decided to test the capabilities of the latest generation of AI language models, a recently created programme called ChatGPT<sup>1</sup> 3.5, released in November 2022. A language model is a computer programme designed to simulate conversation with human users, typically through text-based communication channels such as messaging apps or websites. Language models use natural language processing (NLP) and machine learning algorithms to understand and interpret user inputs and provide appropriate responses. They can help users with a wide range of tasks, including answering questions, providing customer support, scheduling appointments, ordering products, and more. There are two main types of language models: rule-based and AI-powered. Rule-based language models use pre-defined rules and decision trees to determine their responses to user inputs, while AI-powered language models use machine learning algorithms to continuously learn and improve their responses based on user interactions. Language models can be deployed across a wide range of sectors, including academia, e-commerce, healthcare, finance, and customer service. They can help schools automate repetitive tasks, reduce costs, and improve student satisfaction by providing quick and personalised support. Overall, language models represent a rapidly growing area of innovation in the field of artificial intelligence and have the potential to revolutionise the way academia interacts with their students. Apart from the above mentioned, let us list several other pertinent features that convinced us this language model is capable of handling essay assessment (Elkins & Chun, 2020; Gao et al., 2022):

Natural Language Processing (NLP) capabilities: ChatGPT uses state-of-the-art NLP models that enable it to understand and respond to user inputs in a way that feels more human-like. This means that ChatGPT can provide more accurate and relevant responses to teacher queries.

1. Large Knowledge Base: ChatGPT has been trained on a vast amount of text data, making it an excellent source of knowledge. It can answer a wide range of questions and provide useful insights on various essay topics.
2. Continuous Learning: ChatGPT is an AI-powered language model that continuously learns from every interaction it has with users. This means that over time, ChatGPT becomes more accurate and efficient at answering questions and providing support.

3. Versatility: ChatGPT can be integrated into a wide range of platforms and applications, including websites, mobile apps, and messaging platforms. This makes it a versatile solution for teachers looking to provide better student support or automate repetitive tasks.
4. Reliability: ChatGPT is built on the latest technology and is designed to provide reliable and consistent performance. This means that teachers can rely on ChatGPT to provide high-quality support and assistance to their students.
5. Ability to provide instructions and feedback: This is the essence of the essay assessment procedure. High-quality feedback is a time-consuming and highly coveted element of any essay assessment.
6. Ability to challenge and reject incorrect premises: A lot of essays contain factual/logical mistakes, and it is vital that the AI programme recognises and points out those mistakes. For instance, many AI programmes will usually not red-flag sentences of the following type: "Madrid is the capital of France.", since it is a grammatically acceptable sentence, however, still pragmatically unacceptable.
7. Ability to clarify its point of view: Obtaining a "raw" AI opinion is certainly helpful, however, AI's ability to provide further evidence or rationale for why it deems something to be right is a highly desirable feature.
8. Less prone to erroneous answers than the rest of the field: It is worth mentioning we are fully aware that any computer programme of our choosing is still "just" a piece of software that will, from time to time, provide incoherent, erroneous or superfluous answers. However, this language model is one of the best at what it is supposed to do, and it is getting better with constant updates.

Let us briefly mention ChatGPT's limitations and disadvantages (Floridi & Chiriatti, 2020):

1. Lack of emotional intelligence: While ChatGPT can understand the context of the conversation, it does not have affective qualities as humans do. This can lead to limitations in providing empathetic responses.
2. Limited knowledge: Although ChatGPT has access to a vast amount of data and knowledge, it still has limitations in some specific areas of knowledge, and it may not always provide accurate or reliable information.

<sup>1</sup> <https://openai.com/blog/chatgpt/>

3. Inability to understand non-textual input: ChatGPT can only process text-based input and cannot understand other forms of communication like images or sounds. This can lead to limitations in providing comprehensive responses.
4. Potential biases: ChatGPT learns from the data it is trained on, which may contain biases that could be perpetuated in its responses. Developers make constant efforts to reduce biases, nevertheless, they will always remain part of any language model, because there will always be user-induced biases.
5. Limited creativity: Although ChatGPT can generate text and answer questions, it may not be able to provide creative or unique responses that humans can.
6. Limited ability to understand cultural and social nuances: ChatGPT may not be able to understand the cultural and social nuances that exist in human communication, which can lead to limitations in providing appropriate and culture-specific responses.

Fortunately, some of the drawbacks mentioned in the text above are irrelevant to our study (for instance, item no. 3), since the entirety of our data is text-based. Furthermore, the advantages of ChatGPT should outweigh any possible disadvantage since this language model is rapidly getting better and during this research and writing of this paper, it has received several incremental updates<sup>2</sup>, which made this language model far superior to the rest of the field.

#### STUDY DESIGN AND RESEARCH METHODS

Our cross-sectional design study uses concurrent mixed methods (Brewer & Hunter, 2006). We decided to implement this cross-sectional design since it is a type of observational study that involves collecting data from a specific population during a concrete and limited time window (which is the case here), thus making it the most suitable for examining the prevalence of some outcome (Creswell & Creswell, 2018) and making inferences about the characteristics of the population from which we drew our sample (Gray et al., 2007). Concurrent mixed methods refer to our wish to combine both qualitative and quantitative data to comprehensively analyse the task at hand and use the best of both methods to our advantage. Another reason for our mixed-methods approach is hidden in the nature of the research problem. Our research deals with personal/professional opinions of teach-

ers that span both the qualitative and quantitative spectrum (Creswell & Plano Clark, 2007), thus we wanted to employ both research methods.

To obtain valid and generalisable data, we needed to find the minimum number of students that we had to include to have a representative sample. We used the following formula (Sauro & Lewis, 2016):

$$N = K^2 \frac{s^2}{m^2}$$

where the variables are as follows:

*K* is a constant (1.96 for a 95% confidence level or 1.645 for a 90% confidence level – we opted for a higher confidence level to decrease uncertainty in our sample variable),  
*s* is the standard deviation as a proportion of the mean, which is 52% of the mean or 0.52,  
*m* is the desired margin of error, also expressed as a proportion of the mean (we opted for 0.15 corresponding to 15%) (Sauro & Lewis, 2016).

If we plot all these numbers into the formula, we get:

$$N = 1.96^2 \frac{0.52^2}{0.15^2}$$

$$N = 3.8416 \frac{0.2704}{0.022502}$$

$$N = 1.0380544 \times 0.0225$$

$$N = 46.1344$$

Some students were unavailable because they did not sit the exam, did not want to participate because they were abroad or were otherwise unavailable due to professional/personal reasons. By including more students than the statistical minimum we wanted to obtain a more representative sample and more statistically sound and generalisable data. We obtained all the necessary consents, according to Article 2 of the Montenegrin Personal Data Protection Law which prescribes that the processing of data relating to individuals may be carried out for a lawful purpose or with the prior consent of the data subject. Furthermore, personal data may be processed for statistical purposes or the purposes of scientific research, subject to the provision of appropriate safeguards. Our students were informed which personal data would be processed (name, family name and grades), for what purpose (scientific research) and what safeguards were employed (anonymisation, randomisation and point-to-point encryption).

<sup>2</sup> The initial launch of ChatGPT in November 2022 utilised the GPT-3.5 model. However, on March 14, 2023, a GPT-4-based version, which is the latest OpenAI model, was introduced and made accessible to paid subscribers on a restricted basis.

This was followed by the phase where the three teachers would enter all the names and grades in their respective Excel sheets. These Excel sheets were combined into a master Excel sheet where we would assign a unique random ID to each student and then randomly order the students listed on the sheet (using the Excel RAND function). In this manner, the author of this paper was unable to trace any particular essay to any particular student. This measure was taken to avoid even the appearance of bias and increase scientific rigour. Even though it was completely outside of the control of the author of this paper, we were fortunate enough and were able to collect a sample which was roughly gender-balanced (we ended up with 44 females and 34 males).

In terms of data gathering, we chose a computer-based self-administered questionnaire because it is exceptionally easy to deploy (de Leeuw & Hox, 2008) and the data are automatically and conveniently saved in an Excel file. To ensure the validity and reliability of our survey questionnaire, we employed the following measures:

- We had a clear definition of the research objectives and the specific information we aimed to gather through the survey. This helped us ensure that the questionnaire focused on relevant and meaningful questions.
- We conducted a pilot test of the questionnaire with a small sample of respondents to identify any potential issues, such as unclear or confusing questions, ambiguous response options, or formatting problems. We also gathered feedback from the pilot test participants and made necessary revisions to improve the questionnaire.
- We used established and validated scales from the existing research literature. This helped ensure that the items have been tested for reliability and validity in previous studies (Dillman, 2007).
- We avoided using leading or biased questions. We designed questions that are neutral and unbiased, avoiding leading or suggestive language that may influence respondents' answers. Additionally, we used balanced response options and avoided including assumptions or opinions in the questions.
- We considered question order. We arranged the questions in a logical and coherent sequence.

By implementing these measures, hopefully, we enhanced the validity and reliability of our survey questionnaire and improved the quality of the data collected. Our questionnaire was mostly in line with a Likert-type response format. This was done because it is a very convenient and time-saving manner of collecting data. Furthermore, all data is easily retrievable and reanalysable. All respondents were

our fellow teachers, who regularly assess essays as a part of their examination process and whose students belong to the research population. We asked the teachers to enter their anonymous feedback in the form we provided them with. The questionnaire form contained an introductory statement and clear instructions and was organised as a collection of written queries, which contained several clear questions with exhaustive and well-drawn options. These options were a mix of balancing items and a seven-point rating scale, with the last question designed as an open-ended one. While choosing a rating scale, we respected the following three criteria:

- a) Not more than seven points. Cowan (2015, based on Miller, 1956) argued that human beings have a limited capacity to process information and can only reliably make about seven distinct choices or distinctions. This author argues that this limit is related to the capacity of our working memory, which is the system that temporarily holds and manipulates information.
- b) Provision of a middle alternative in a scale. It is good to include a middle alternative because it serves as a reference point for the respondents who truly belong in the middle of the scale. Research shows that if no middle alternative is present, respondents tend to randomly choose other options which has a detrimental effect on data validity (O'Muircheartaigh et al., 2000).
- c) Assign labelled words to the options. Respondents can react quickly and answer with more confidence if the provided options are clearly labelled and easily distinguishable (Brace, 2004). The five and seven options provided below are very easy to distinguish, especially for our fellow teachers, since most of these options are also used as letter grades, something our colleagues are very familiar with.

These three criteria were used for our questionnaire, which contained eight close-ended questions and one open-ended question. This mix of questions was used to make the most of what close-ended and open-ended questions have to offer (Peterson, 2000), with the aim of eliciting the following information:

#### **1. What is your overall opinion about the essay?**

Options: a) Very positive b) Positive c) Neutral d) Negative e) Very negative

#### **2. How did you grade the following elements:**

a. The level of critical analysis and the quality of judgement

Options: a) Outstanding b) Excellent c) Very good d) Good e) Satisfactory f) Sufficient g) Fail

b. The quality of arguments and line of reasoning  
Options: a) Outstanding b) Excellent c) Very good  
d) Good e) Satisfactory f) Sufficient g) Fail

c. The originality of the essay i.e., personal touch  
vs. generic answers, lateral thinking  
Options: a) Outstanding b) Excellent c) Very good  
d) Good e) Satisfactory f) Sufficient g) Fail

d. Cohesion and coherence  
Options: a) Outstanding b) Excellent c) Very good  
d) Good e) Satisfactory f) Sufficient g) Fail

e. Grammar elements  
Options: a) Outstanding b) Excellent c) Very good  
d) Good e) Satisfactory f) Sufficient g) Fail

f. Technical aspects  
Options: a) Outstanding b) Excellent c) Very good  
d) Good e) Satisfactory f) Sufficient g) Fail

### 3. What is the overall grade?

Options: a) Outstanding b) Excellent c) Very good  
d) Good e) Satisfactory f) Sufficient g) Fail

### 4. Provide additional reasons for such a grade.

Options: this is an open-ended question, for everything we missed with our close-ended questions.

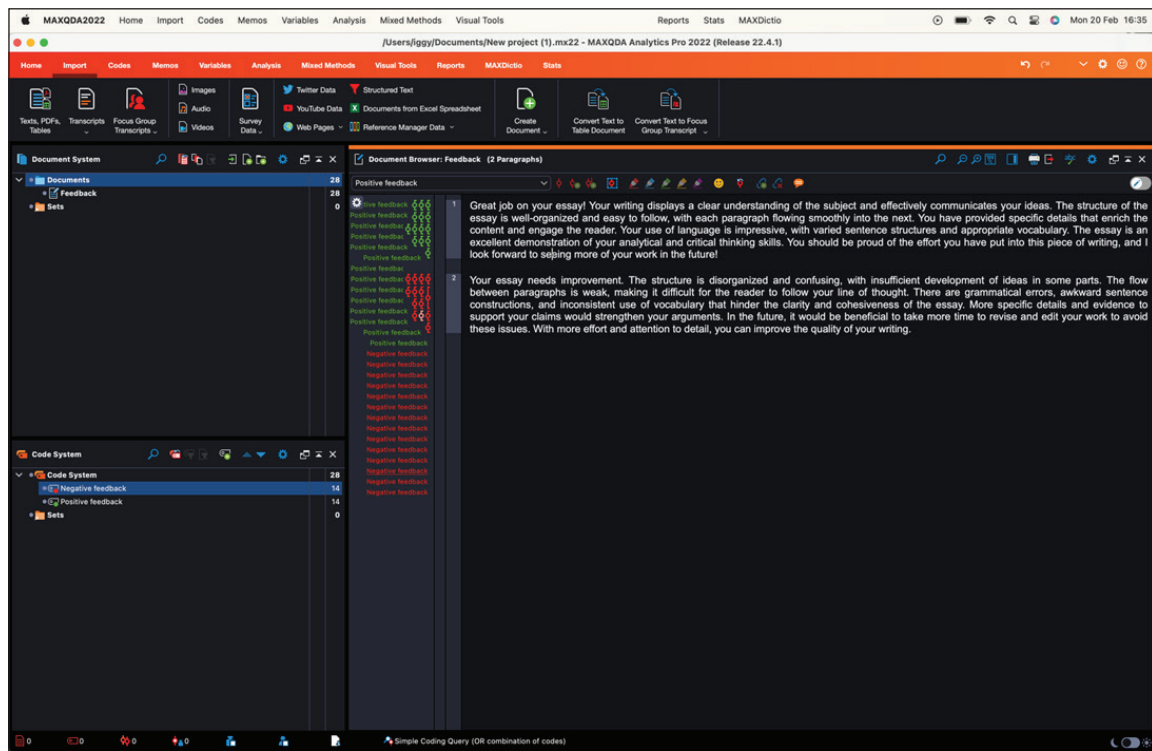
The questions are organised along the lines recommended by Creswell (2006). First, we have an ice-breaker question, followed by seven close-ended questions and finishing with an open-ended question. Additionally, this question order was deliberately chosen to capture the respondents' attention by offering them to first elaborate on their opinion, which should be affectively engaging for the respondents. We tried to word our questions in the most concise way possible and to avoid over-elaboration. We opted for the seven-point rating scale because it is generally aligned with the grading scale of our university and would be instantly recognisable for the respondents, which would facilitate and expedite the completion of the questionnaire. Another reason for this is that a limited range of clearly defined answers would be easily comparable with the answers provided by the AI language model. On the other side, the open-ended question enabled us to obtain high-value answers that provided uniquely good insight, a greater understanding of the nuances pertaining to the essay assessment process and helped build a better picture of why teachers graded the way they did (Davis et al., 2005). Furthermore, this open-ended question enabled us to capture any

additional information missed by the closed-ended questions. Here, the respondents are free from any undue influence of a predetermined nature of close-ended questions. With this combination of having both open and close-ended questions, we wanted to strike a balance between the possibility of receiving answers irrelevant to our research and the possibility of limiting the respondents' frame of reference too much. The questionnaire form was uploaded to a cloud service, the link was provided to the respondents, and they had two months (December 2022 & January 2023 – the winter examination diet) to complete the task. In this manner, we wanted to remove almost all pressure and allow our colleagues to provide quality and meaningful answers at their convenience.

## CODING AND DATA ANALYSIS

As soon as the respondents' answers started trickling in, we were immediately ready to start working with ChatGPT and observe the output we would get. The author would copy-paste the essay into the ChatGPT interface and ask it the same questions that we asked our respondents. To make sure ChatGPT "understood" our questions and provided appropriate answers, we would ask it one question at a time. When the author copy-pasted the relevant essay, ChatGPT would<sup>3</sup> first pre-process the inserted text to remove any extraneous information, such as headers, footers, and references. This ensured that the text was clean and ready for analysis. The essay was then tokenised into individual sentences. This allowed ChatGPT to analyse the text more granularly and identify patterns in sentence structure and language use. The essay was analysed for various aspects of language use, including grammar, vocabulary, and syntax. This helped to identify errors or issues with the language use that could impact the overall quality of the writing. The essay was evaluated for the quality of its content, including the strength of arguments, the use of evidence, and the relevance of the content to the topic at hand. This analysis was based on the patterns that ChatGPT had learned from a large dataset of essays, intending to identify factors that are associated with high-quality writing. The essay was evaluated for its coherence and organisation, including the use of topic sentences, transitions, and logical connexions between ideas. This helped to ensure that the essay was well-structured and easy to understand. Finally, based on the results of the previous analyses, ChatGPT would assign an overall evaluation score to the essay. This score was based on the patterns that had been learned

3 <https://platform.openai.com/docs/model-index-for-researchers>



**Figure 1:** MAXQDA Analytics Pro 2022 interface.

from the training data and it reflected the quality of the writing, as judged by ChatGPT's algorithms. Chat GPT would generate an answer and the author would copy the answer to a Word document thanks to the "Copy for ChatGPT" Chrome extension since copying from ChatGPT was not enabled by default at the time. Once the answer was in a .docx file format, it was ready to be thematically and contextually analysed. Two of our more experienced colleagues were tasked with analysing the received data independently and then jointly to negotiate and iron out the differences. For them to be able to perform a proper analysis, they needed to analyse the data elicited from the close-ended questions and code for the open-ended questions. For close-ended questions, they decided to simply count the frequency of different instances of answers. Analysing this type of closed-ended questions in linguistic research typically involves assigning numerical values to the responses (Tuzzi, 2001). After defining the response options, the coders assigned numerical values to each response option ranging from (5) for Outstanding to (0) for Fail. In the meantime, we created a coding sheet to record the responses of each participant. The coding sheet consisted of a column for the participant's ID number and a separate column for each closed-ended question as well as for the open-ended one.

For open-ended questions, coding was done with the help of MAXQDA Analytics Pro 2022 (Figure 1) along the lines of thematic content analysis and answer categorisation. The content analysis had the aim of discovering themes and subthemes until reaching the saturation point (Kiger & Varpio, 2020). A theme is a construct that is repetitive and conceptually links different expressions into a meaningful group (Ryan & Bernard, 2003). Since we had two coders, we needed to implement measures which would ensure inter-coder reliability (ICR). ICR refers to the consistency and agreement between multiple coders or researchers who independently analyse and code qualitative data. To increase and ensure inter-coder reliability, we implemented the following measures (Kurasaki, 2000):

- Development of clear and detailed coding guidelines that provided explicit definitions and examples of the codes to be used. The guidelines covered the coding process, code definitions, inclusion and exclusion criteria, and decision rules for resolving coding ambiguities.
- Organisation of thorough training sessions to familiarise coders with the coding guidelines and ensure a shared understanding of the coding process. We used practice datasets or

pilot coding sessions to calibrate coders and address any discrepancies or uncertainties before they began coding the actual data.

- Establishment of consistent coding procedures to be followed by both coders. This included maintaining uniformity in coding software, naming conventions and file organisation to avoid any discrepancies due to procedural differences.
- Scheduling regular meetings or discussions among coders to address questions, clarify coding ambiguities, and share insights.
- Conducting periodic checks to compare and assess coding consistency between coders. This involved comparing a subset of overlapping coded data or conducting a reliability check on a random sample of coded data.

Based on the analysed data, we identified five main themes that were the most prevalent in the 78 analysed essays:

1. Academic motivation: this theme focuses on students' motivations for learning and their attitudes towards academic pursuits. This theme was relatively prevalent in the students' essays exploring students' intrinsic or extrinsic motivation, their passion for a particular subject, their goals and aspirations, or their perception of the value and relevance of their studies. Within this theme, students elaborated and expressed their opinions about intrinsic motivation factors like curiosity and students' innate desire to explore, discover, and learn new things out of pure curiosity. They also mentioned enjoyment (or the lack thereof) and pleasure in studying a particular topic or subject together with the perception that the subject matter was (not) personally meaningful, applicable to their lives, or aligned with their interests and goals. This theme also covers essays mentioning extrinsic motivation factors such as grades and academic performance and the students' desire to achieve high grades, recognition, or academic accolades such as external rewards or incentives, such as certificates, scholarships, or tangible benefits, that motivate students to engage in learning activities. Moreover, there was students' motivation to outperform others or meet certain standards set by their peers or educational institutions coupled with external pressures or expectations from parents, teachers, or society to excel academically.
2. Learning strategies and study habits: this theme originated from the essays in which students described the approaches they employed to

engage with their studies. It included examining their study habits, time management skills, note-taking methods, problem-solving techniques, or strategies they utilised to enhance comprehension and retention of the material. Students elaborated on their study habits, such as setting specific study times, creating a dedicated study environment, or using study aids like mnemonic devices. A certain number of students would break down their study material into manageable chunks, establish study goals, and utilise active learning techniques (e.g., summarising, self-testing) to enhance comprehension and retention. Around two-thirds of students used different organisation tools and techniques, such as planners, calendars, or digital apps, to help students manage deadlines, assignments, and study materials effectively. Some 20% of all students reported developing various note-taking methods and techniques, such as summarising key points, creating concept maps, using bullet points, or annotating texts.

3. Perceptions and viewpoints: a certain number of analysed essays from our corpus dealt with students' perspectives, attitudes, emotions, and thoughts related to the topics they were writing about. This was our favourite part of the research since the students expressed a rich tapestry of diverse perspectives, reflecting unique backgrounds, experiences, and viewpoints. Excitement and fascination were common emotions expressed by students who found the subject matter intriguing, stimulating their curiosity and motivating deeper exploration. Scepticism or critical questioning emerged as emotions expressed by students who engaged in a thoughtful examination of the topics, challenging assumptions, or seeking evidence to support their viewpoints. Some students experienced frustration or uncertainty when grappling with complex ideas or when their perspectives were confronted with contradictory viewpoints. The variety of emotions expressed demonstrated the complex interplay between affective factors and students' learning experiences, highlighting the impact of emotional engagement on motivation, understanding, and critical thinking. The essays showcased a wide range of attitudes towards the topics, demonstrating that students approached them from various angles and had differing opinions or beliefs. Additionally, students' personal experiences, cultural backgrounds, and prior knowledge strongly influenced their interpretations and engagement with the subject matter, leading to the formation of distinct viewpoints.

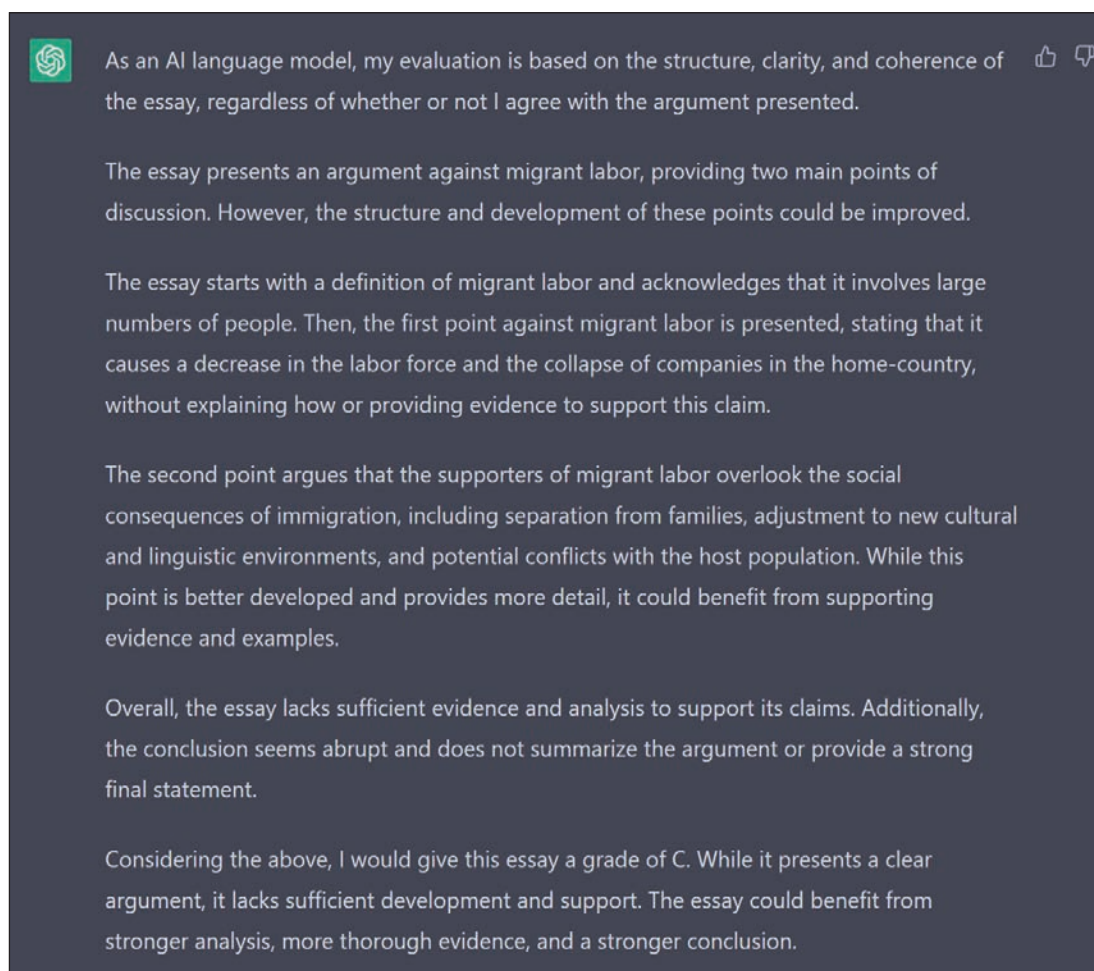
4. Personal growth and development: this theme explored the personal and intellectual growth students experience throughout their academic journey. These essays encompassed changes in students' knowledge, skills, self-confidence, critical thinking abilities, worldview, or understanding of themselves and their potential. The thematic analysis findings indicate that students reported acquiring new knowledge and skills throughout their academic journey. This growth was reflected in a diverse range of subject areas and domains. Research results indicate that approximately 85% of the students reported an expansion of their knowledge base and demonstrated an increased understanding of the subject matter, including mastering content-specific information, gaining expertise in research methodologies, or developing proficiency in practical skills relevant to their field of study. The research findings suggest that students experienced an increase in self-confidence as they progressed in their academic journey. Approximately 75% of the students expressed a greater sense of self-assurance and belief in their abilities to tackle academic challenges. Around 60% of the students reported actively reflecting on their learning experiences, identifying areas for improvement, and setting meaningful goals to enhance their academic performance.
5. Socio-cultural themes: these essays encompassed a wide range of interesting social topics like diversity, inclusion, cultural identity, social norms, gender roles, or how students' backgrounds and experiences shape their academic engagement and viewpoints. The research findings suggest that students' essays explored the importance of diversity and inclusion within the academic context. Our research results indicate that approximately 80% of the essays reflected discussions on the significance of embracing diverse perspectives, cultures, and backgrounds. Our students highlighted the benefits of diverse learning environments, fostering empathy, and promoting a sense of belonging among students from various backgrounds. Around 70% of the essays expressed students' reflections on how their cultural identity shapes their experiences and perspectives in the academic setting. Approximately 75% of the essays highlighted the recognition of societal expectations, such as academic achievement, gender roles, or conformity, and their impact on students' experiences and choices. Research results show that around 60% of the essays reflected discussions on gender equality, stereotypes, and the importance of creating equitable opportunities for all students.

## RESULTS AND DISCUSSION

The analysis of our data aimed at answering the question pertaining to the viability of using ChatGPT as a teachers' digital assistant. In practical terms, this can be measured by the level of congruity between the overall grade provided by teachers and by ChatGPT. It stands to reason that the more congruous these two camps are, the more useful ChatGPT is for the teachers. If the opinions of the two groups diverge significantly, then teachers may not have that many benefits from using AI as their essay assessment companion because their assessment strategies and viewpoints would be incompatible.

Simultaneously with the analysis of the data obtained from our colleagues, we were "feeding" ChatGPT with the essay sent by our teachers. We asked ChatGPT to analyse the provided essays along with the same questions given to the teachers. From time to time, ChatGPT would provide a nonsensical or vague answer, but through the option Regenerate answer we "insisted" on providing a clear opinion on the matter at hand. It is worth mentioning that both the respondents and ChatGPT had the same assessment criteria that they needed to consider because when assessing an essay written by students for whom English is a second language, it is important to consider certain criteria that acknowledge their language proficiency and specific needs. Here are the key criteria:

1. Assess the students' ability to effectively communicate ideas and concepts in English. Consider factors such as grammar, vocabulary usage, sentence structure, and overall language fluency. Keep in mind that minor errors may be expected in second-language writing, so focus on comprehensibility and clarity rather than nitpicking every mistake.
2. Evaluate the students' comprehension and grasp of the topic or subject matter. Assess whether they have clearly understood the key concepts, theories, or ideas related to the essay prompt. Look for evidence of critical thinking, analysis, and the ability to support arguments with relevant examples or evidence.
3. Assess the students' ability to structure their essays logically and coherently. Look for a clear introduction, body paragraphs that develop and support the main ideas, and a conclusion that effectively summarises the key points. Consider the use of cohesive devices to ensure smooth transitions between paragraphs and ideas.
4. Evaluate the students' ability to present and develop a well-reasoned argument or perspective. Assess their use of evidence, logical reasoning, and critical analysis to support their claims. Look



**Figure 2:** ChatGPT's assessment of an essay, rendered in under 5 seconds (sample).

- for the students' ability to evaluate different viewpoints, anticipate counterarguments, and provide persuasive explanations.
5. Consider the students' understanding of cultural nuances and the appropriate use of language in a particular context. Assess their ability to effectively communicate their ideas while considering cultural differences, ensuring that their writing is respectful and inclusive.
  6. Evaluate whether the students have fully addressed the essay prompt or question. Assess whether they have met the requirements and objectives of the assignment, including any specific guidelines or criteria provided.
  7. Consider the students' ability to bring fresh perspectives or unique insights to the topic. Assess their creativity in presenting ideas, engaging the reader, or using innovative approaches to communicate their thoughts effectively.

Remember to provide constructive feedback that helps the students improve their language skills and academic writing. Offer specific suggestions for improvement in language usage, organisation, or critical thinking, highlighting strengths and areas that need development. It can also be helpful to provide additional support or resources tailored to second language learners to help them further enhance their English proficiency.

Based on these criteria ChatGPT would provide the output (sample) as featured on Figure 2.

After eliciting all the required answers from ChatGPT, we were ready to compare these two sets of data and see the level of agreement between the teachers and ChatGPT. For the close-ended questions, it was a simple matter of comparing which options were selected by teachers and which by ChatGPT. One of the best ways of measuring the level of agreement between the two sets of such data was for us to calculate the intraclass correlation coefficient (ICC) for

consistency between raters. ICC for consistency is a statistical measure used to assess the degree of agreement or consistency among multiple raters or observers when measuring a continuous outcome variable. It quantifies the proportion of the total variance in the measurements that is attributable to the true differences between the subjects or items being rated.

ICC for consistency is typically used when the raters are measuring the same subjects or items on the same scale or using the same measurement technique. It provides an estimate of the reliability or agreement among the raters' scores. ICC values range from 0 to 1, where:

ICC = 0 indicates no agreement or consistency among the raters' scores.

ICC = 1 indicates perfect agreement or consistency among the raters' scores.

Higher ICC values indicate greater agreement or consistency among the raters' scores, suggesting that the variability in the measurements is mostly due to true differences between the subjects or

items rather than measurement error.

To have a valid ICC calculation and data interpretation, we adopted the following criteria:

*As a rule of thumb, researchers should try to obtain at least 30 heterogeneous samples and involve at least 3 raters whenever possible when conducting a reliability study. Under such conditions, we suggest that ICC values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability and values greater than 0.90 indicate excellent reliability.* (Koo & Li, 2016)

Our research included 78 highly heterogeneous samples, since the data vary among the samples (Wallis et al., 2014), with 3+1 (human + AI) raters. Thus, according to Koo & Li (2016), our research method is more than valid for the calculation of the ICC. The following table provides all the inputs necessary for the calculation of the ICC.

**Table 1: Grades given by Rater 1.**

Student ID	Rater 1	#20	2	#40	2	#60	4
#1	3	#21	5	#41	4	#61	4
#2	1	#22	3	#42	2	#62	3
#3	2	#23	1	#43	4	#63	4
#4	3	#24	4	#44	5	#64	4
#5	2.5	#25	2	#45	1	#65	2
#6	2	#26	5	#46	3	#66	2
#7	4	#27	0	#47	2	#67	5
#8	4	#28	2.5	#48	2	#68	3
#9	2.5	#29	5	#49	5	#69	5
#10	0	#30	4	#50	3	#70	5
#11	2.5	#31	4	#51	3	#71	4
#12	5	#32	3	#52	5	#72	2
#13	2	#33	4	#53	2	#73	5
#14	5	#34	0	#54	3	#74	3
#15	4	#35	5	#55	2	#75	3
#16	1	#36	3	#56	4	#76	3
#17	3	#37	3	#57	3	#77	2
#18	5	#38	2	#58	4	#78	3
#19	5	#39	5	#59	0		

**Table 2: Grades given by Rater 2.**

Student ID	Rater 2	#20		#40		#60	
#1	2	#21	5	#41	3	#61	3
#2	0	#22	2	#42	1	#62	4
#3	1	#23	2.5	#43	3	#63	3
#4	3	#24	3	#44	5	#64	3
#5	3	#25	1	#45	1	#65	1
#6	1	#26	5	#46	4	#66	1
#7	2.5	#27	0	#47	1	#67	5
#8	3	#28	4	#48	1	#68	2
#9	4	#29	5	#49	4	#69	5
#10	0	#30	3	#50	2	#70	5
#11	4	#31	3	#51	2	#71	3
#12	5	#32	4	#52	4	#72	1
#13	3	#33	3	#53	1	#73	5
#14	5	#34	0	#54	4	#74	4
#15	3	#35	5	#55	1	#75	4
#16	1	#36	4	#56	3	#76	2
#17	4	#37	4	#57	4	#77	2.5
#18	5	#38	1	#58	3	#78	3
#19	5	#39	5	#59	0		

**Table 3: Grades given by Rater 3.**

Student ID	Rater 3	#20		#40		#60	
#1	4	#21	5	#41	5	#61	5
#2	2.5	#22	4	#42	3	#62	2
#3	3	#23	0	#43	5	#63	5
#4	2	#24	5	#44	5	#64	5
#5	4	#25	3	#45	0	#65	3
#6	3	#26	5	#46	2	#66	3
#7	5	#27	2.5	#47	3	#67	4
#8	5	#28	3	#48	3	#68	4
#9	3	#29	5	#49	4	#69	5
#10	0	#30	5	#50	4	#70	5
#11	2	#31	5	#51	4	#71	5
#12	5	#32	2	#52	5	#72	3
#13	2	#33	5	#53	3	#73	5
#14	5	#34	0	#54	2	#74	2
#15	5	#35	5	#55	3	#75	2
#16	2.5	#36	2	#56	5	#76	1
#17	2.5	#37	2	#57	2	#77	2
#18	5	#38	3	#58	5	#78	3
#19	5	#39	5	#59	0		

**Table 4: Grades given by ChatGPT.**

Student ID	ChatGPT	#20	2	#40	2	#60	4
#1	2	#21	5	#41	5	#61	4
#2	1	#22	3	#42	2	#62	4
#3	3	#23	1	#43	3	#63	4
#4	3	#24	4	#44	3	#64	4
#5	3	#25	4	#45	2	#65	3
#6	2	#26	4	#46	3	#66	2
#7	5	#27	1	#47	3	#67	5
#8	5	#28	3	#48	3	#68	3
#9	3	#29	4	#49	5	#69	5
#10	1	#30	4	#50	3	#70	5
#11	3	#31	4	#51	3	#71	5
#12	5	#32	3	#52	5	#72	3
#13	2	#33	3	#53	2	#73	2
#14	5	#34	1	#54	2	#74	3
#15	5	#35	5	#55	2	#75	4
#16	1	#36	4	#56	4	#76	3
#17	3	#37	4	#57	2	#77	3
#18	3	#38	2	#58	4	#78	3
#19	4	#39	5	#59	1		

**Table 5: Two-factor ANOVA inputs used to calculate the ICC.**

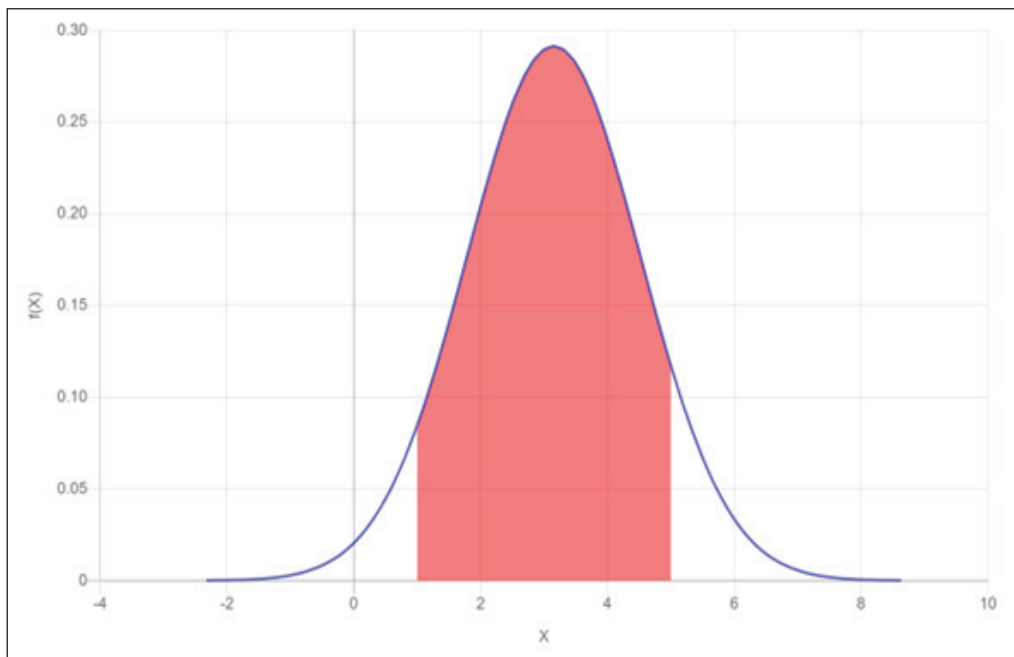
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Rows	236.3397	77	3.069347	9.7718	3.55E-20	1.458228
Columns	0.314103	1	0.314103	1	0.320443	3.965094
Error	24.1859	77	0.314103			
Total	260.8397	155				
<b>ICC</b>	<b>0.81433</b>					

Therefore, based on the calculated ICC value, we can safely conclude that there is good interrater reliability based on the two sets of data. The ICC above 0.8 suggests a high degree of agreement or similarity among the observations in terms of the following:

- ICC values above 0.8 indicate a high degree of consistency or reliability among the observations. This suggests that the measurement tool

or procedure used in the study is producing consistent and reliable results.

- ICC values above 0.8 suggest that there is low variability among the observations, meaning that the raters are producing consistent results across the range of values being assessed.
- ICC values above 0.8 also indicate a strong correlation among the observations. This means



**Chart 1: Grade distribution curve – teachers.**

that the raters are producing results that are highly correlated with each other, suggesting that they have similar opinions towards the same underlying construct.

Generally speaking, ChatGPT was more lenient and would award higher grades to students when compared to human counterparts. This was, at least partially, due to the nature of ChatGPT's algorithm. As an artificial intelligence language model, ChatGPT does not have emotions, biases, or subjective opinions that can influence its evaluation of essays. It relies on its training data and algorithms to analyse and score essays based on predefined criteria such as grammar, syntax, coherence, and relevance to the prompt. In some cases, ChatGPT may be more lenient when assessing essays compared to human evaluators because it is designed to be more forgiving of minor errors or deviations from standard writing conventions. For example, it may overlook minor grammatical mistakes or unconventional phrasing that a human evaluator would notice and count against the essay. The second reason for higher leniency is that we did not set any additional or stricter parameters in terms of essay assessment. We used default criteria and let ChatGPT behave and assess "as is" within the given key criteria mentioned in the text above. On that same note, another interesting takeaway from the analysis of the presented data is the fact that ChatGPT did

not fail any student. The reason for this (according to the AI itself) is that this language model is capable of crunching and interrelating a huge amount of data, so there was always at least some redeeming factor which prevented it from failing a student. Truth be told, students at the Faculty of Science and Mathematics consistently show a high level of knowledge, so there is not much room for talking about failing anybody.

Regarding the distribution of grades, both distributions follow the Gaussian distribution, which makes their mean the centre of their probability distribution. If we input the parameters obtained for the distribution of grades given by the teachers and the ones given by the AI language model, we can see there are a lot of similarities. The following are the population mean ( $\mu$ ) and population standard deviation ( $\sigma$ ) provided for teachers:

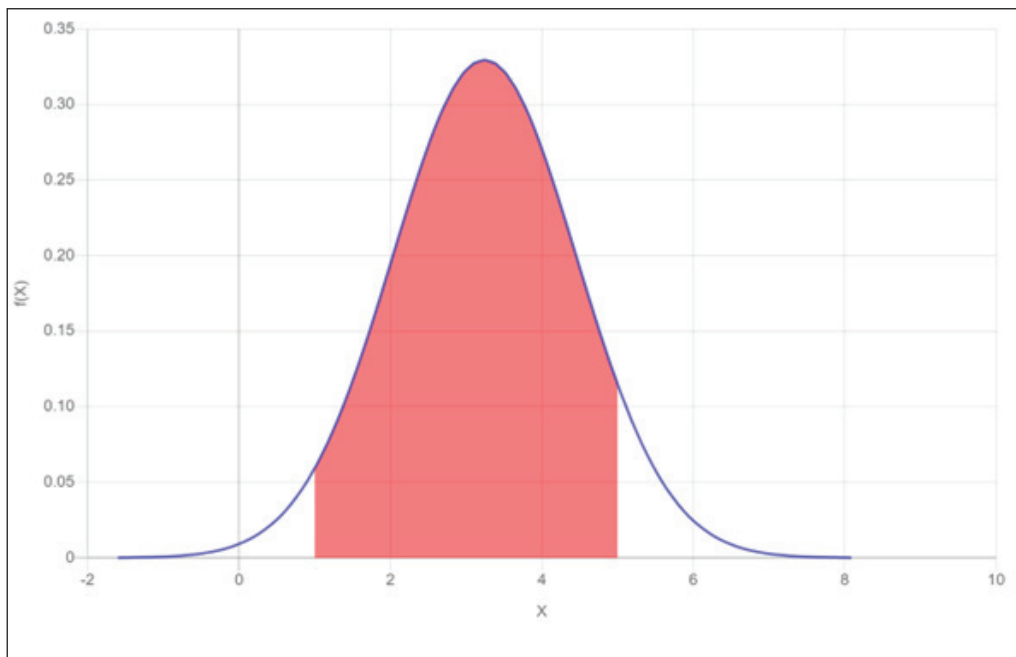
Population Mean ( $\mu$ ) = 3.15

Population Standard Deviation ( $\sigma$ ) = 1.37

We have the following graph that corresponds to the specified event  $0 \leq X \leq 5$ :

Normal Distribution:  $\Pr(0 < X < 5) = 0.85$

The given data describes the grade distribution curve for teachers, which follows a Gaussian distribution. The mean (population mean) of this distribution is 3.15, and the standard deviation (population standard deviation) is 1.37. Based on this information, we can analyse the probability



**Chart 2: Grade distribution curve – ChatGPT.**

of a certain event occurring within this distribution. In this case, the event is specified as  $0 \leq X \leq 5$ , where  $X$  represents the grade. To calculate the probability of this event, we look at the corresponding area under the normal distribution curve. The probability is given as  $\Pr(0 < X < 5) = 0.85$ . This probability indicates that there is an 85.35% chance of a randomly selected grade falling between 0 and 5 within the teacher's grade distribution. This suggests that a significant majority of grades awarded by the teachers lie within this range, as indicated by the high probability value.

With ChatGPT, the following are the population mean ( $\mu$ ) and population standard deviation ( $\sigma$ ):

Population Mean ( $\mu$ ) = 3.24

Population Standard Deviation ( $\sigma$ ) = 1.21

We have the following graph that corresponds to the specified event  $0 \leq X \leq 5$ :

Normal Distribution:  $\Pr(0 < X < 5) = 0.89$

In the data provided, the population mean ( $\mu$ ) is 3.24, and the population standard deviation ( $\sigma$ ) is 1.21. This indicates the characteristics of a grade distribution, similar to the previous example. The probability of the event  $0 \leq X \leq 5$ , as calculated from the normal distribution graph, is  $\Pr(0 < X < 5) = 0.89$ . This probability implies an 89.46% chance of a randomly selected grade

falling within the range of 0 to 5. When comparing this result with the previous entry, we can observe some slight differences. The probability obtained in this case (0.89) is slightly higher than the probability in the previous example (0.85). This indicates that, based on the data, there is a greater likelihood of receiving a grade within the specified range when it comes to ChatGPT. Furthermore, the population mean ( $\mu$ ) in this case (3.24) is slightly higher than the population mean in the previous example (3.15). This suggests a shift towards higher grades or a generally higher average grade within the grade distribution. Similarly, the population standard deviation ( $\sigma$ ) in this case (1.21) is smaller than the standard deviation in the previous example (1.36). A smaller standard deviation indicates less variability in the grades and a narrower spread around the mean. Overall, these comparisons show that the data suggests a slightly higher average grade and a tighter distribution compared to the previous data. These similarities are correlated to the calculated ICC which statistically proves ChatGPT is a viable and useful essay grading assistant. The only slight difference that is very difficult to discern from these two graphs is that teachers tend to have their grades more spread out (higher standard deviation), whereas ChatGPT tends to be more conservative, thus making the peak of the second graph slightly more pronounced.

## CONCLUSIONS AND CONSIDERATIONS

As an AI language model, ChatGPT can assist teachers/professors/assistants in assessing essays in numerous and very practical ways. One of the most useful ways of assistance is related to automated grading and feedback. Professors can use ChatGPT to provide instant feedback to students, reducing the need for professors to spend hours and hours assessing numerous essays. The importance of this time-saving aspect cannot be overstated. AI can drastically reduce the time necessary to assess students' essays. The average time for ChatGPT to render an essay assessment was consistently below 30 seconds, while the teachers reported they needed 5 to 15 minutes to thoroughly read an essay and an additional 5 to 15 minutes to grade it. Thus, an essay assessment can take anywhere between 10 and 30 minutes per essay. Of course, assessment time depends greatly on the length of the essay and the complexity of the topic, just to name a few. What is even more impressive, ChatGPT can analyse the essays based on predefined criteria. Thus, professors' feedback can be tailored to their needs and target specific essay deficiencies which need to be corrected. What has positively surprised us, was the level of detail that GPT was able to extract from the essays and how granular its analysis was. Each analysis was very detailed with a sound and logical foundation. Even though 78 essays have many things in common, we felt ChatGPT paid individual attention to each essay, rather than resorting to some generic statements with little value in terms of deeper insight. To sum it all up, based on our study and the data it yielded, an AI-assisted essay assessment is the right way of embracing new technologies and using them in such a way that they can help teachers by speeding up the tedious side of essay assessment and making the whole process more enjoyable and ultimately more beneficial for the students. On a more generalised plane AI-assisted assessment, teaching and/or learning may be a new paradigm shift and bring about the didactic and methodological renewal of our education (Kukanja Gabrijelčič, 2015). Our research has shown that ChatGPT is a reliable essay assessment tool, especially when it comes to analysing large amounts of data in a short amount of time. ChatGPT can assess essays at a much faster rate than humans can. It can evaluate essays in a matter of seconds, while it may take humans hours or even days to read and evaluate many essays. While assessing those essays, ChatGPT assesses them objectively, based on the patterns it has learned from the training

data it was trained on. When ChatGPT assesses an essay, just like humans, it analyses various aspects of the text, including the coherence of ideas, the strength of arguments, the use of evidence, the organisation and structure of the text, and the quality of language and grammar.

Another useful feature of ChatGPT is plagiarism detection. With the advent of the internet, the way students approach their essay writing tasks has been changed forever. Plagiarism is now a fact of life and is here to stay, however, under those circumstances, professors should have effective plagiarism detection software at their disposal to avoid blatant and more subtle copying. ChatGPT can assist with plagiarism detection by providing an additional resource for comparing and analysing texts. You can input a suspicious text into ChatGPT and ask it to compare it with known sources or specific passages. ChatGPT can provide insights into similarities or differences between the texts and help identify potential instances of plagiarism. ChatGPT can help you identify if a text has been paraphrased from a source by generating alternative versions or rephrasing sentences. By comparing the suspicious text with the paraphrased versions, you can assess the level of similarity and potential plagiarism. ChatGPT can be integrated with existing plagiarism detection tools or algorithms. It can assist in automating the process by generating queries or conducting searches using specific phrases from the suspicious text, helping to identify potential matches or similarities in external sources. ChatGPT can provide contextual understanding and analysis of the suspicious text, comparing it with known sources to identify inconsistencies or discrepancies. It can highlight potential passages that need further investigation for plagiarism. It's important to note that while ChatGPT can be a helpful tool in plagiarism detection, it should not be solely relied upon. It's still crucial to use established plagiarism detection tools, manual comparison, and critical analysis to ensure accuracy and thoroughness in identifying plagiarism. ChatGPT can complement these existing methods and provide an additional perspective during the process.

To reach its full potential and become even more useful for teachers, ChatGPT needs to be integrated into teachers' work. Integrating AI tools into teachers' work in a supportive manner involves considering the specific needs and context of teachers and their students. AI tools can provide comprehensive training and professional development programmes for teachers to familiarise them with AI tools, their capabilities, and their limitations. Teachers should receive

guidance on how to effectively use AI tools to enhance their teaching practices, streamline administrative tasks, and support student learning. Furthermore, we have to define clear pedagogical objectives for integrating AI tools. We must identify specific areas where AI can provide value, such as personalised learning, formative assessment, or administrative tasks. Moreover, we need to align the use of AI tools with the goals of the curriculum and teaching methodologies to ensure they enhance the teaching and learning experience. After this, we should involve teachers in the development and selection of AI tools. Engage them in collaborative discussions, seek their feedback, and encourage the co-creation of AI-based solutions that address their unique needs. This collaboration ensures that the tools are relevant, practical, and tailored to the specific teaching context. AI tools have to be introduced gradually and incrementally, allowing teachers to gain familiarity and confidence. We could start with pilot projects or small-scale implementations, gather feedback, and refine the integration based on teacher and student experiences. Gradual implementation minimises disruption and allows for continuous improvement. Institutions must ensure that the necessary infrastructure, technical support, and resources are in place to support the integration of AI tools. By considering these factors and providing ongoing support, training, and collaboration opportunities, teachers can effectively integrate AI tools into their work in a manner that enhances their teaching practices and benefits student learning.

However, it is important to note that ChatGPT as well as our study come with some limitations. When using AI for essay grading, there are potential biases that can arise. If the AI model is trained on a dataset that is biased or unrepresentative, it may learn and perpetuate those biases in its grading. For example, if the training data is predominantly from a specific demographic or cultural background, the AI may not be as accurate or fair when grading essays from different backgrounds. These models can be biased towards certain types of language or writing styles. This can result in essays that deviate from the model's preferred style being graded lower, even if the content is well-constructed and insightful. AI models may be biased towards certain topics or perspectives, leading to potential inconsistencies or unfairness in grading. For example, if an AI model is trained on essays primarily focused on Western literature, it may not adequately evaluate essays on non-Western literature. Additionally, while ChatGPT can

analyse text quickly, it may not be able to capture the nuances of language and meaning that a human reader could readily identify. This is especially true for complex or creative writing, where the writer may use figurative language or other devices that are not easily detected by a machine. The only real grudge that any evaluator who is interested in using ChatGPT as his/her evaluation assistant may hold against this remarkable piece of an AI language model is when it fails to detect what we call "spaced-apart incongruities". If a student would make an initial claim that A led to B and B led to C and then, later, that A and C are utterly unrelated and if these two claims were separated by a body of text longer than several sentences, ChatGPT sometimes struggled to make connexion and detect these incongruities. However, it is only fair to mention that in a sizeable number of cases (83%) ChatGPT performed admirably well and indeed detected internal inconsistencies, which is more than we can ask from a newly developed software.

In terms of the limitations of our study, a higher participation percentage would certainly increase our confidence interval and produce more reliable data. The same could be said for the length of our study since it encompasses the essays written during one semester. Inclusion of more semesters, i.e., longer study would yield more reliable and generalisable data.

Future research can focus on addressing the limitations of AI in understanding complex linguistic devices and inconsistencies. Experts will inevitably develop more sophisticated linguistic models that capture the nuances and complexities of language. This includes exploring advanced natural language processing (NLP) techniques, such as semantic parsing, discourse analysis, and understanding figurative language. These models should be able to comprehend and interpret linguistic devices, such as metaphors, sarcasm, irony, and rhetorical strategies. This will enhance AI models' ability to understand and interpret language in context. Contextual understanding involves considering the broader meaning, background knowledge, and cultural references associated with the text. This can be achieved through leveraging contextual embeddings, incorporating world knowledge databases, and building models that can reason and infer based on the given context. All stakeholders are interested in exploring techniques for training AI models with limited annotated data. Many complex linguistic devices and inconsistencies are context-specific and require fine-grained training data. Techniques such as transfer learning, few-shot learning, and active

learning can be explored to overcome data limitations and improve AI's ability to handle linguistic complexities. Also, we must develop methods to make AI models more transparent and explainable in their decision-making processes. This includes techniques for interpreting the model's internal representations, generating explanations for predictions, and providing feedback on the model's understanding of linguistic devices and inconsistencies. Explainability helps users, such as teachers and students, trust and understand the AI's judgments. Experts are already investigating the integration of multiple modalities, such as text, images, and audio, to enhance AI's under-

standing of linguistic devices and inconsistencies. By incorporating visual or auditory cues along with textual information, AI models can gain a more comprehensive understanding of the context and improve their ability to interpret complex linguistic features. As with all research, we must not forget about the ethical implications of AI in language understanding and inconsistency handling. This includes studying potential biases, fairness issues, and unintended consequences that may arise when AI models are used in real-world applications. Thus, we have to develop guidelines and frameworks to ensure the responsible and ethical use of AI in language-related tasks.

## ALI LAHKO OCENJEVANJE ESEJEV S POMOČJO UMETNE INTELIGENCE REŠI PROFESORJE NAVZKRIŽNA RAZISKAVA Z UPORABO MEŠANIH METOD IZVEDENA NA UNIVERZI V ČRNI GORI

Igor IVANOVIĆ

Univerza Črne Gore, Filološka fakulteta, Danila Bojovića bb, 81400 Nikšić, Črna Gora  
e-mail: iggybosnia@ucg.ac.me

### POVZETEK

*Ta raziskava je proučevala sposobnost ChatGPT-a, jezikovnega modela, ki temelji na umetni inteligenci, za ocenjevanje študentskih esejev. Rezultate ocenjevanja ChatGPT-a smo primerjali in statistično analizirali glede na rezultate, ki smo jih prejeli od naših kolegov profesorjev. Korpus naše analize je sestavljalo 78 študentskih esejev, rezultati analize pa so pokazali, da obstaja visoka stopnja skladnosti med ChatGPT in ocenjevalci, kar potrjuje visoka vrednost korelacijskega koeficienta znotraj razreda (ICC). Ta ICC nakazuje, da je ChatGPT lahko v veliko pomoč profesorjem pri ocenjevanju študentskih pisnih nalog. Statistično gledano so porazdelitve ocen vseh ocenjevalcev sledile Gaussovi porazdelitvi. Posebej pomembno je poudariti, da sta bili povprečni vrednosti za ocenjevalce in ChatGPT 3,15 oziroma 3,24, z majhnimi razlikami v njihovih standardnih odstopanjih. Porazdelitve verjetnosti so pokazale tudi subtilne razlike, pri čemer ima ChatGPT nekoliko višjo povprečno oceno in nižjo stopnjo variabilnosti. To pomeni, da je bil ChatGPT nekoliko bolj dosleden pri svojih ocenah v primerjavi z našimi kolegi ocenjevalci. Poleg naštetega je dodatna prednost ChatGPT njegova hitrost, ki znaša približno 30 sekund na esej, v primerjavi z ocenjevalci, ki jim je bilo potrebno med 10 do 30 minut. Vendar smo ugotovili, da je bil ta jezikovni model nekoliko popustljiv, verjetno zaradi svoje objektivne algoritemske narave. Edini resnični očitek v zvezi s ChatGPT (različica 3.5) je, da občasno ni uspel prepoznati določenih nedoslednosti v esejih, vendar je bila tudi ta težava večinoma odpravljena z izdajo različice 4.0. Da bi bila orodja, ki temeljijo na UI, kot je ChatGPT, optimalno uporabljena v izobraževalnem okolju, naša študija priporoča postopno integracijo, ki vključuje temeljito usposabljanje učiteljev, jasne pedagoške cilje in kakovostno informacijsko infrastrukturo za podporo.*

**Ključne besede:** ChatGPT, samodejno ocenjevanje, jezikovni modeli umetne inteligence, ocenjevanje esejev, povratne informacije o esej, merjenje ocenjevanja, obdelava naravnega jezika

## SOURCES AND BIBLIOGRAPHY

- Brace, Ian (2004):** Questionnaire Design: How to Plan, Structure and Write Survey Material for Effective Market Research. London, Kogan Page.
- Brewer, John & Albert Hunter (2006):** Foundations of Multimethod Research: Synthesizing Styles. Thousand Oaks, Sage.
- Cowan, Nelson (2015):** George Miller's Magical Number of Immediate Memory in Retrospect: Observations on the Faltering Progression of Science. *Psychological Review*, 122, 3, 536–541.
- Creswell, John (2006):** Qualitative Inquiry and Research Design: Choosing among Five Approaches. London, Sage Publications.
- Creswell, John & David Creswell (2018):** Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Los Angeles, Sage.
- Creswell, John & Vicki Plano Clark (2007):** Designing and Conducting Mixed Methods Research. Thousand Oaks, Sage.
- Davis, James, Smith, Tom & Peter Marsden (2005):** General Social Surveys, 1972–2004 Cumulative File. ICPSR04295-v1. Chicago, National Opinion Research Centre.
- Dehouche, Nassim (2021):** Plagiarism in the Age of Massive Generative Pre-trained Transformers (GPT-3). *Ethics in Science and Environmental Politics*, 21, 17–23.
- de Leeuw, Edith & Joop Hox (2008):** Self-Administered Questionnaires. In: de Leeuw, Edith, Hox, Joop & Don A. Dillman (eds.): *International Handbook of Survey Methodology*. Abingdon, Routledge, Routledge Handbooks Online.
- Dexter, Emily (1935):** The Effect of Fatigue or Boredom on Teachers' Marks. *The Journal of Educational Research*, 28, 9, 664–667.
- Dillman, Don (2007):** Mail and Internet surveys: The tailored design method (2<sup>nd</sup> ed.). New York, John Wiley.
- Elkins, Katherine & Jon Chun (2020):** Can GPT-3 pass a Writer's Turing test? *Journal of Cultural Analytics*, 5, 2, 1–16.
- Elpidorou, Andreas (2022):** Boredom and Cognitive Engagement: A Functional Theory of Boredom. *Review of Philosophy and Psychology*, forthcoming.
- Erturk, Sinan, van Tilburg, Wijnand & Eric Igou (2022):** Off the Mark: Repetitive Marking Undermines Essay Evaluations due to Boredom. *Motivation and Emotion*, 46, 264–275.
- Floridi, Lucioano & Massimo Chiriatti (2020):** GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30, 4, 681–694.
- Gao, Catherine, Howard, Frederick, Markov, Nikolay, Dyer, Emma, Ramesh, Siddi, Luo, Yuan & Alexander Pearson (2022):** Comparing Scientific Abstracts Generated by ChatGPT to Original Abstracts Using an Artificial Intelligence Output Detector, Plagiarism Detector, and Blinded Human Reviewers. *NPJ Digital Medicine*, 6, 1.
- Gao, Jianmin (2021):** Exploring the Feedback Quality of an Automated Writing Evaluation System Pigai. *International Journal of Emerging Technologies in Learning (ijET)*, 16, 11, 322–330.
- Gray, Paul, Williamson, John, Karp, David & Jean Charles Dalphin (2007):** The Research Imagination: An Introduction to Qualitative and Quantitative Methods. Cambridge, Cambridge University Press.
- Grillon, Christian, Quispe-Escudero, David, Mathur, Ambika & Monique Ernst (2015):** Mental fatigue impairs emotion regulation. *Emotion (Washington, D.C.)*, 15, 3, 383–389.
- Kiger, Michelle & Lara Varpio (2020):** Thematic Analysis of Qualitative Data: AMEE Guide No. 131. Medical Teacher.
- Koo, Terry & Mae Li (2016):** A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15, 2, 155–163.
- Kukanja Gabrijelčič, Mojca (2015):** Student-tailored Textbook? International Comparative Analysis of Questions and Tasks in History Textbooks. *Annales, Series Historia et Sociologia*, 25, 2, 385–398.
- Kurasaki, Karen (2000):** Intercoder Reliability for Validating Conclusions Drawn from Open-ended Interview Data. *Field Methods*, 12, 179–194.
- Marcora, Samuele, Staiano, Walter & Victoria Manning (2009):** Mental Fatigue Impairs Physical Performance in Humans. *Journal of Applied Physiology*, 106, 3, 857–864.
- Miller, George (1956):** The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review*, 63, 81–97.
- Mizuno, Kei, Tanaka, Masaka, Yamaguti, Kouzi, Kajimoto, Osami, Kuratsune, Hirohito & Yasuyoshi Watanabe (2011):** Mental Fatigue Caused by Prolonged Cognitive Load Associated with Sympathetic Hyperactivity. *Behavioural and Brain Functions*, 7, 17.
- O'Muircheartaigh, Colm, Krosnick, Jon & Armen Helic (2000):** Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data. Harris School of Public Policy Studies, University of Chicago, Working Papers.
- Peterson, Robert (2000):** Constructing Effective Questionnaires. Thousand Oaks, Sage.
- Ryan, Gery & Russell H. Bernard (2003):** Techniques to Identify Themes. *Field Methods*, 15, 1, 85–109.
- Sauro, Jeff & James Lewis (2016):** Quantifying the User Experience: Practical Statistics for User Research. Amsterdam, Elsevier.
- Tuzzi, Arjuna (2001):** Subjects on Using Open and Closed-Ended Questions. In: Borra, Simone, Rocci, Roberto, Vichi, Maurizio & Martin Schader (eds.): *Advances in Classification and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin – Heidelberg, Springer.
- Wallis, Allen, Roberts, Harry & George Shultz (2014):** The Nature of Statistics (Illustrated). Mineola, Dover Publications.