מיהו המעריך המדויק ביותר? הלימה בין ציוני ChatGPT ועמיתים לבין ציוני המרצה עבור פרויקטי סטודנטים ברמות איכות שונות

Montathar Faraon
Kristianstad University
Emailmontathar.faraon@hkr.se

מאיה אושר
HIT מכון טכנולוגי חולון
mayau@hit.ac.il

Who Is the Most Accurate Evaluator? Alignment of ChatGPT and peer grades with instructor grades across varying levels of student project quality

Maya Usher
HIT Holon Institute of Technology
mayau@hit.ac.il

Montathar Faraon
Kristianstad University
montathar.faraon@hkr.se

Abstract

The integration of generative artificial intelligence (GenAI) tools into assessment processes in higher education raises critical questions about their alignment with human evaluations. The effectiveness of AI chatbots such as ChatGPT in generating assessments comparable to human evaluators remains unclear, with limited research offering direct comparisons in educational contexts. This quantitative study aims to examine grading alignment across three evaluators of student group projects (ChatGPT, peers, and the course instructor) and to determine whether evaluator agreement varies by project quality. The study involved 184 undergraduate students who submitted a group project and provided peer assessments of their classmates' work. The projects were categorized into three quality levels: low, medium, and high. The analyses revealed that alignment with instructor grading varied systematically by both grading source and project quality. ChatGPT struggled to identify weaker projects and tended to assign them inflated grades, whereas students were better at identifying weaker work but were overly strict toward high-quality projects. In addition, alignment with the instructor's grades was not consistent but depended on project quality, with larger discrepancies observed for ChatGPT's evaluations relative to those of peers. Alignment between ChatGPT's grades and the instructor improved as project quality increased, whereas peer–instructor alignment was strongest for lower-quality work. These findings support a cautious integration of ChatGPT into assessment processes, with the final decision remaining with informed and critical human judgment.

Keywords: ChatGPT, Generative artificial intelligence (GenAI), Higher education, Peer assessment, Peer feedback.

תקציר

שילובם של כלי בינה מלאכותית יוצרת בתהליכי הערכה בהשכלה הגבוהה מעלה שאלות קריטיות בנוגע לאופן ההלימה עם הערכות אנושיות. יעילותם של צ'טבוטים דוגמת ChatGPT בהערכת תוצרי למידה טרם נבחנה דייה וקיים מחקר מוגבל המציע השוואות ישירות בין הערכה זו לבין הערכה אנושית. מטרת מחקר כמותי זה הינה לבחון את מידת ההלימה בין הציונים שניתנו על ידי ChatGPT ועל ידי עמיתים עם ציוני מרצת קורס, תוך הבחנה בין פרויקטי סטודנטים ברמות איכות שונות. במחקר השתתפו 184 סטונדטים שהגישו פרויקט קבוצתי וסיפקו הערכות עמיתים לפרויקטים של חבריהם לכיתה. הפרויקטים קוטלגו לשלוש רמות לפי איכותם: נמוכה, בינונית וגבוהה. ממצאי המחקר הראו כי שני מקורות ההערכה הציגו דפוסי רגישות שונים לרמת איכות הפרויקט המוערך: ChatGPT התקשה לזהות עבודות חלשות ונטה להטות את ציוניהן כלפי מעלה, בעוד שסטודנטים הצטיינו בזיהוי עבודות חלשות, אך החמירו יתר על המידה כלפי עבודות באיכות גבוהה. בנוסף, נמצא כי מידת ההלימה לציוני המרצה אינה עקבית אלא תלויית איכות פרויקט, כאשר פערים גבוהים יותר נמצאו בהערכת ChatGPT אל מול העמיתים. ההלימה בין ציוני ChatGPT למרצה השתפרה ככל שעלתה איכות הפרויקט, בעוד שההלימה בין ציוני העמיתים למרצה היתה גבוהה בעיקר בעבודות באיכות הנמוכה. הממצאים תומכים בשילוב מתון וזהיר של ChatGPT בתהליכי הערכה, כאשר חשוב להותיר את ההכרעה הסופית לשיפוט אנושי מושכל וביקורתי.

מילות מפתח: בינה מלאכותית יוצרת (GenAI), הערכה מבוססת צ'טבוט, הערכת עמיתים, השכלה גבוהה, צ'טג'פט (ChatGPT)

מבוא

שילובה ההולך וגובר של בינה מלאכותית יוצרת (GenAI) במסגרות חינוכיות פתחה אפיקים חדשים לשיטות הערכה חדשניות (Chan & Hu, 2023; Tam, 2024; Usher, 2025). בין היישומים המבטיחים ביותר של GenAI בהקשר של הערכה הוא השימוש בצ'טבוטים דוגמת ChatGPT, שהם מעין מתווכים אוטומטיים הפועלים באמצעות מודלי שפה גדולים (LLMs), ומסוגלים לנהל שיחות משמעותיות ומודעות הקשר (Essel et al., 2022; Labadze et al., 2023). יכולות אלו הופכות אותם לכלי מבטיח לאוטומציה של משימות שונות הקשורות להערכה בסביבות חינוכיות (Okonkwo & Ade-Ibijola, 2021). מכאן, יישום מרכזי ומבטיח בשילוב כלי GenAI הינו הערכה אוטומטית של עבודות סטודנטים, לרבות מתן ציונים וכתיבת משוב (Chan & Hu, 2023; Okonkwo & Ade-Ibijola, 2021). עם זאת, שילוב זה מעלה לא מעט חששות בנוגע לתוקפן של הערכות מבוססות GenAI, שכן ייתכן ואלו אינן מתיישבות עם מטרות תוכנית הלימודים, סטנדרטים דיסציפלינריים וניואנסים הקשריים (Morris et al., 2024; Usher et al., 2025; Venter et al., 2024).

המחקר אודות מידת ההלימה בין הערכה מבוססת GenAI לבין הערכות אנושיות הניב עד כה ממצאים מעורבים. מספר מחקרים דיווחו על הסכמה גבוהה בין ציונים שניתנו על ידי LLMs לבין ציונים מגורם אנושי, במיוחד במשימות מובנות עם מחוון מוגדר מראש וקריטריונים אובייקטיביים (Haudek & Zhai, 2023; Morris et al., 2024). מקורות נוספים הציגו עקביות בינונית-גבוהה בין ChatGPT לבין הערכת מומחים אנושיים (Lu et al., 2024; Pinto et al., 2023). עם זאת, קיימים מחקרים אשר זיהו פערים ניכרים בין שני סוגי ההערכות הללו, בעיקר במשימות פתוחות או מורכבות. למשל, מחקר שנערך לאחרונה זיהה כי ChatGPT העניק בעקביות ציונים גבוהים יותר מאלה של מרצים עבור פרויקטי סטודנטים, עם מתאמים בינוניים בלבד (Usher, 2025), בעוד שמחקר אחר הבחין כי ChatGPT נמנע מנתינת ציונים קיצוניים מטה או מעלה, והפגין דיוק נמוך יותר בשאלות שהיו קשורות הדוקות לתוכן הרצאות הקורס (Flodén, 2025). ממצאים מנוגדים אלו מעוררים שאלות רבות בנוגע למידת המהימנות והתוקף של הערכה מבוססת GenAI, במיוחד סביב תוצרים הדורשים שיפוט ביקורתי, רגישות דיסציפלינרית והבנה הקשרית (Labadze et al., 2023; Tam, 2024; Usher, 2025).

לצד העלייה בשילוב הערכות מבוססות GenAI, הערכת עמיתים נותרה גישה פדגוגית מבוססת בהשכלה הגבוהה, המוכרת כמעודדת מעורבות סטודנטים, חשיבה ביקורתית ולמידה רפלקטיבית (Ocampo et al., 2024; Topping, 2025; Usher & Barak, 2018). כאשר זו נתמכת במחוון מדויק ובהכשרה מתאימה, הערכת עמיתים יכולה להגיע לרמות מהימנות גבוהות ולעיתים אף להשתוות להערכת מרצים (Double et al., 2020; Li et al., 2020). מחקרים מצביעים על כך שמידת ההלימה בין הערכות עמיתים להערכות מרצה עשויה לעלות משמעותית בהתאם למוטיבציה של הסטודנטים, לניסיונם הקודם בהערכה ולפרשנותם את קריטריוני ההערכה (Falchikov & Goldfinch, 2000; Suñol et al., 2016).

שכבה נוספת של מורכבות נוגעת לאופן בו ההלימה עם ציוני המרצה עשויה להשתנות בהתאם לאיכות העבודה המוערכת. ממצאים מוקדמים הציעו כי כלי GenAI עשויים לתפקד באופן שונה בעת הערכת תוצרים באיכות גבוהה לעומת אלו באיכות נמוכה. לדוגמה, מחקר משנת 2025 מצא כי מודלי שפה גדולים זיהו וקיבצו ביעילות רבה יותר תשובות סטודנטים שהיו מלכתחילה באיכות גבוהה, אך התקשו להבחין בין תשובות באיכות נמוכה (Gurin et al., 2025). לעומת זאת, במחקר אחר לא נמצא קשר מובהק בין איכות חיבור של סטודנטים לבין המשוב שנוצר בידי ChatGPT או עמיתים – דבר המצביע על מגבלות אפשריות ברגישות לרמת הביצוע בשני המקורות (Banihashem et al., 2024). אי-עקביות זו מדגישה את הצורך לבחון לא רק את מידת ההלימה הכוללת בין הציונים המוענקים על ידי מקורות שונים, אלא גם כיצד זו עשויה להשתנות בהתאם לרמות שונות של איכות עבודות סטודנטים.

מטרה ושאלת המחקר

מטרת המחקר הנוכחי הינה לבחון את מידת ההלימה בין ציונים שניתנו לפרויקטי סטודנטים על ידי ChatGPT ועל ידי עמיתים עם ציוני מרצת קורס, תוך הבחנה בין פרויקטי סטודנטים ברמות איכות שונות (נמוכה, בינונית, גבוהה). מחקר זה מונחה על ידי שתי שאלות מחקר:

1. באיזו מידה הציונים שניתנו לפרויקטי סטודנטים על ידי ChatGPT ועל ידי העמיתים נמצאים בהלימה עם ציוני מרצת הקורס?
2. האם ובאיזה אופן מידת ההלימה עם ציוני מרצת הקורס משתנה בהתאם לאיכות הפרויקט המוערך?

אוכלוסיית וסביבת המחקר

במחקר השתתפו 184 סטודנטים לתואר ראשון (147 נשים ו- 37 גברים) אשר לקחו חלק בקורס חובה העוסק במדידה והערכה בשנים האקדמיות 24-2023 ו- 25-2024. כחלק מדרישות הקורס, הסטודנטים עבדו על פרויקט בקבוצות של 3-4 משתתפים, אשר חולק לשלושה שלבים: פיתוח מערך מחקר מבוסס שאלון, הערכת עמיתים, והערכה מבוססת ChatGPT. ראשית, הסטודנטים עבדו בקבוצות על פיתוח שאלון מקוון בנושא לבחירתם, שכלל לפחות ארבע שאלות סגורות ושתי שאלות פתוחות. בשלב השני, כל סטודנט ביצע הערכת עמיתים לשני פרויקטים של חבריו לכיתה. התהליך היה אנונימי, והסטודנטים השתמשו במחוון מפורט שחולק לששת חלקי הפרויקט: מטרת המחקר, אוכלוסיית המחקר, שאלות המחקר, שאלות סגורות בשאלון, שאלות פתוחות בשאלון והקדמה לשאלון. עבור כל אחד מששת חלקי המחוון צוינו הנחיות מילוליות וכן ציון כמותי מתוך הציון הכולל. הסטודנטים העניקו ציונים מספריים ומשוב כתוב עבור כל אחד מששת החלקים. לבסוף, הסטודנטים ביצעו הערכה מבוססת ChatGPT, בה התבקשו להפיק הערכה שהתבססה על אותו המחוון בכדי לספק ציונים מספריים ומשוב כתוב עבור כל אחד מששת חלקי הפרויקט. הסטודנטים נשענו על הנחיות (פרומפטים) שניתנו להם מראש על ידי המרצה, אך התבקשו להמשיך ולנהל שיח מתמשך מול ChatGPT ככל שיהיה בכך צורך.

שיטת המחקר, כלי המחקר וניתוח

המחקר התבסס על מתודולוגיה כמותית, בה נאספו ונבחנו הציונים שניתנו לפרויקטי הסטודנטים על ידי שלושת מקורות ההערכה: ChatGPT, עמיתים ומרצת הקורס. הנתונים שנאספו כללו את הציונים שניתנו לכל פרויקט על ידי המקורות השונים, ונערכה ביניהם השוואה לצורך זיהוי פערים ודפוסי הלימה בינם לבין ציוני מרצת הקורס. ציוני המרצה היוו מדד בסיס לסיווג הפרויקטים לפי איכותם; החלוקה נעשתה בהסתמך על היסטוגרמה וחישוב האחוזונים ה-33 וה-67 של כלל המדגם. בהתאם לתוצאות האחוזונים, הפרויקטים סווגו לשלוש רמות איכות:

- איכות נמוכה (ציון מרצה ≥ 80, 64 סטודנטים)
- איכות בינונית (ציון 81–86, 65 סטודנטים)
- איכות גבוהה (ציון ≤ 87, 55 סטודנטים)

הנתונים נותחו באמצעות סטטיסטיקה תיאורית, כולל חישוב ממוצעים וסטיות תקן, ולאחר מכן ניתוח שונות חד־כיווני במדידות חוזרות (Repeated Measures ANOVA) עם תיקון Greenhouse–Geisser והשוואות פוסט־הוק מסוג Bonferroni. בנוסף, נבחנו המתאמים בין כל זוג מקורות הערכה באמצעות מתאם פירסון, תוך הבחנה בין פרויקטים ברמות איכות שונות. הפער בין ציוני המעריכים לבין ציונים המרצה, בהתאם לשלוש רמות איכות הפרויקטים, נותחו באמצעות ניתוח שונות חד־כיווני נוסף, וחושבו הבדלי ממוצעים (*Mdiff*) עבור כל מעריך חלופי. כאשר נמצאו אפקטים מובהקים, בוצעו מבחני פוסט־הוק מסוג Tukey's HSD בכדי לזהות הבדלים ספציפיים בין רמות איכות הפרויקטים.

ממצאים

הלימה בין ציוני ChatGPT וציוני עמיתים לבין ציוני המרצה

על מנת לתת מענה לשאלת המחקר הראשונה, חושבו והושוו הציונים הממוצעים וסטיות התקן שניתנו לפרויקטים מצד כל מקור הערכה. הניתוח חשף דפוס ברור: הציונים הממוצעים שהופקו על ידי ChatGPT היו באופן עקבי גבוהים יותר (M = 91.46, SD = 5.13) בהשוואה לציוני העמיתים ולציוני המרצה שהיו נמוכים יותר אך קרובים זה לזה (M = 85.56, SD = 6.06; M = 83.13, SD = 6.39). ניתוח ANOVA הצביע על הבדל מובהק סטטיסטית בין ציוני שלוש הקבוצות עם גודל אפקט גבוה (η^2 = .42; p < .001, F(1.91, 349.23) = 133.40).

בהמשך, נבחנה מידת ההלימה בין ציוניהם של שלושת המעריכים (ראו טבלה 1). נמצא מתאם חיובי בינוני בין ציוני העמיתים עם ציוני המרצה (p < .001, r(182) = .48). מנגד, המתאם בין ציוני ChatGPT לבין ציוני המרצה היה מתון יותר (p < .01, r(182) = .24). בין ציוני ChatGPT לבין ציוני העמיתים לא נמצא מתאם מובהק כלל (p = .48, r(182) = .05).

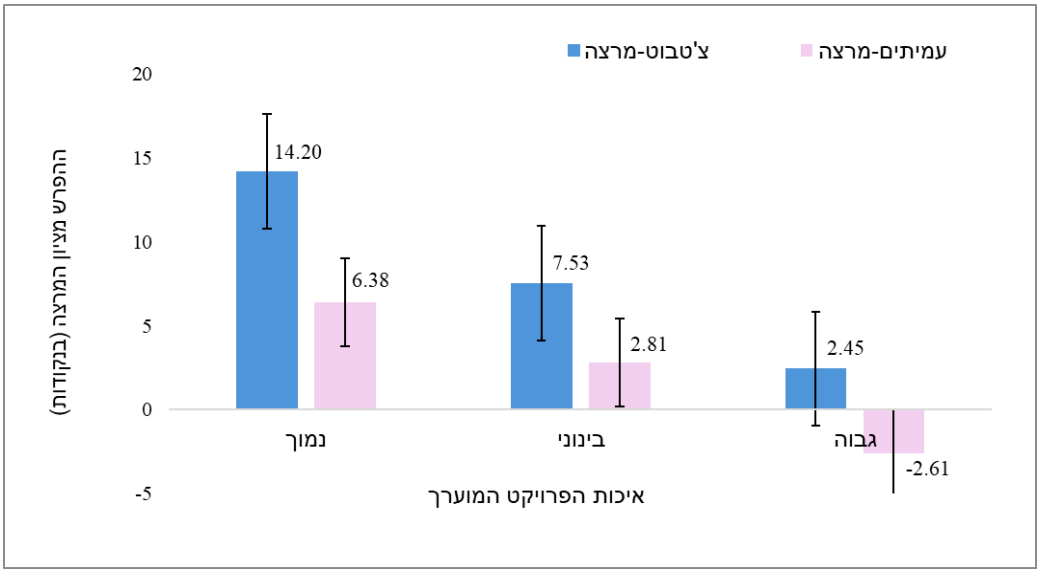
טבלה 1. מתאמים עבור ציוני ChatGPT , העמיתים ומרצת הקורס

מקור הערכה	ChatGPT	עמיתים	מרצה
ChatGPT	1		
עמיתים	0.05	1	
מרצה	0.24**	0.48***	1

הערה: N = 184. ** p < .01; *** p < .001

שונות בהלימת הציונים בהתאם לאיכות הפרויקט המוערך

על מנת לתת מענה לשאלת המחקר השנייה, נבחנו והושוו הפערים בין ציוני ChatGPT או העמיתים לבין ציוני המרצה, בהתאם לשלוש רמות איכות הפרויקטים (נמוכה, בינונית, גבוהה). איור 1 מציג את ההפרש בין ממוצעי הציונים שניתנו על ידיChatGPT או העמיתים לבין ציוני המרצה, בהתאם לרמת איכות הפרויקט.



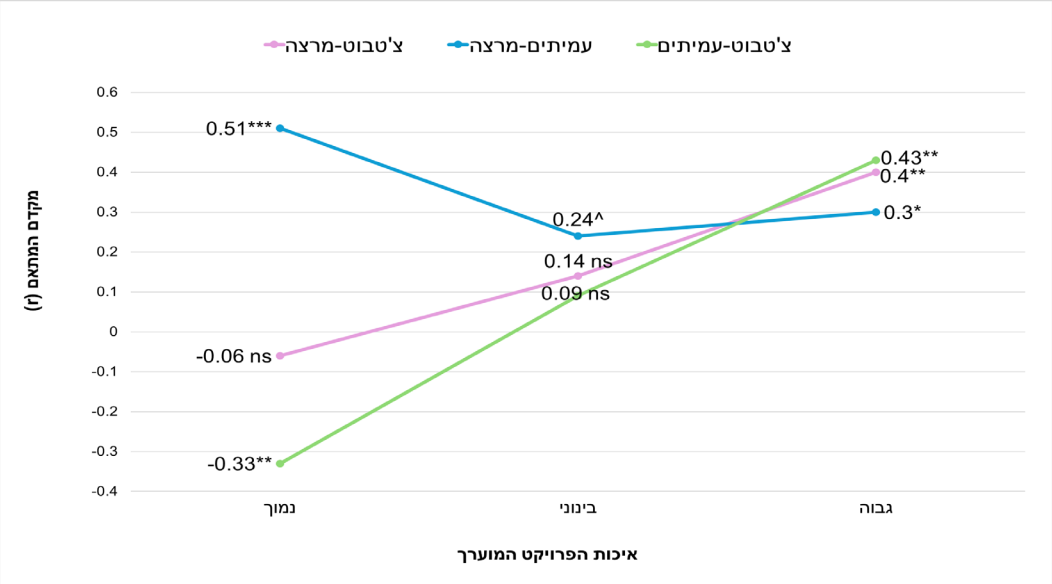
איור 1. הפרש הממוצעים בין ציוני ChatGPT וציוני עמיתים לציוני המרצה, לפי רמות איכות הפרויקטים.

הניתוח חשף השפעה מובהקת של רמת איכות הפרויקט על הפער שבין ציוני ChatGPT לציוני המרצה (p < .001, F(2, 181) = 71.01), עם גודל אפקט גבוה (η^2 partial = .44). בממוצע, ChatGPT העניק ציונים גבוהים

באופן משמעותי מהמרצה ברמת האיכות הנמוכה (M_{diff} = 14.20 points, SD = 5.89), גבוהים באופן מתון יותר ברמת האיכות הבינונית (M_{diff} = 7.53 points, SD = 4.92) וגבוהים רק במעט ברמת האיכות הגבוהה (*Mdiff* = 2.45 points, SD = 5.36). השוואות פוסט-הוק אישרו כי הפערים בקבוצת הפרויקטים באיכות הנמוכה היו גבוהים באופן מובהק מאלו בקבוצת האיכות הבינונית (M_{diff} = 6.67 points, p < .001) ובקבוצת האיכות הגבוהה (M_{diff} = 11.75 points, p < .001). בנוסף, נמצא כי הפערים בקבוצת הפרויקטים באיכות הבינונית היו גבוהים באופן מובהק מאלו בקבוצת האיכות הגבוהה (M_{diff} = 5.08 points, p < .001).

דפוס דומה, אך מתון יותר, נמצא גם בקרב הערכות העמיתים. הניתוח חשף השפעה מובהקת של רמת איכות הפרויקט על הפער שבין ציוני העמיתים לציוני המרצה (F(2, 181) = 43.22, p < .001), עם גודל אפקט גבוה (partial η^2 = .32). בממוצע, העמיתים העניקו ציונים גבוהים מהמרצה ברמת האיכות הנמוכה (*Mdiff* = 6.38 points, SD = 6.09) וברמה הבינונית (M_{diff} = 2.81 points, SD = 5.49), אך ציונים נמוכים מעט מהמרצה בקבוצת האיכות הגבוהה (M_{diff} = -2.61 points, SD = 3.76). השוואות פוסט-הוק הצביעו על כך שהפערים בקבוצת הפרויקטים באיכות הנמוכה היו גבוהים באופן מובהק מאלו בקבוצת האיכות הבינונית (M_{diff} = 3.57 points, p < .001) ובקבוצת האיכות הגבוהה (M_{diff} = 8.99 points, p < .001). בנוסף, נמצא כי הפערים בין הפרויקטים בקבוצת האיכות הבינונית והגבוהה היו מובהקים סטטיסטית (M_{diff} = 5.42 points, p < .001).

להשלמת הממצאים, חושבו מתאמי פירסון בנפרד עבור כל אחת משלוש רמות איכות הפרויקטים, על מנת לבחון את מידת ההלימה בין ציוני המרצה לבין אלו שניתנו על ידי ChatGPT או העמיתים (ראו איור 2). בקבוצת הפרויקטים *באיכות הנמוכה*, ציוני העמיתים הראו מתאם בינוני ומובהק סטטיסטית עם ציוני המרצה (r = .51, p < .001), בעוד שציוני ChatGPT לא הראו כל מתאם מובהק בקבוצת איכות זו (r = -.06, p = .62). בקבוצת הפרויקטים *באיכות הבינונית*, מידת ההלימה נותרה נמוכה בקרב שני מקורות ההערכה: ציוני העמיתים הראו מתאם נמוך וגבולי עם ציוני המרצה (r = .24, p = .05) וציוני ChatGPT לא הראו מתאם מובהק כלל (r = .14, p = .26). דפוס זה התהפך בקבוצת הפרויקטים *באיכות* הגבוהה: ציוני ChatGPT היו במתאם חזק יותר עם ציוני המרצה (r = .40, p = .001) מאשר ציוני העמיתים (r = .30, p = .03). דפוסים אלה מצביעים על כך שמידת ההלימה בין ציוני ChatGPT לציוני המרצה משתפרת ככל שעולה איכות הפרויקט, בעוד שהערכת העמיתים הייתה מדויקת יותר בעיקר בפרויקטים באיכות הנמוכה.



איור 2. מתאמי פירסון עבור כל אחת משלוש רמות איכות הפרויקט המוערך

דיון ומסקנות

ממצאי המחקר מספקים תובנות לגבי מידת הדיוק וההתאמה בין הערכה מבוססת ChatGPT לבין מעריכים אנושיים עבור פרויקטים קבוצתיים של סטודנטים ברמות איכות שונות. נראה כי מידת ההלימה של הציונים שניתנו על ידי ChatGPT או העמיתים, עם הציונים שניתנו על ידי מרצת הקורס, אינה עקבית אלא משתנה בהתאם לזהות המעריך ולאיכות הפרויקט המוערך. הממצאים הדגישו הבדלים ברורים ומובהקים בין מקורות ההערכה השונים מבחינת האופן

בו הם מעניקים ציונים לפרויקטים באיכויות שונות. ChatGPT העניק ציונים גבוהים יותר באופן עקבי בהשוואה לעמיתים ולמרצה – תופעה המלמדת על נטייה להערכת־יתר (grade inflation). ממצא זה מאשש עדויות קודמות בנוגע לנטיית ChatGPT להערכה מקלה (Flodén, 2025; Usher, 2025), אך מנוגד למחקרים אחרים שדיווחו על הלימה גבוהה בין GenAI לבני אדם (Lu et al., 2024; Morris et al., 2024). המתאם המתון בלבד בין ציוני ChatGPT לציוני המרצה עשוי להעיד על היגיון הערכתי שונה – כזה המדגיש עמידה ביעדים טכניים על פני שיפוט איכותי רגיש או בעל הקשר. ממצא זה תואם מחקרים המצביעים על מגבלות GenAI בזיהוי ניואנסים דיסציפלינריים ובהבנה הקשרית (Morris et al., 2024; Usher, 2025; Venter et al., 2024). מנגד, המתאם הבינוני והמובהק שנמצא בין ציוני העמיתים לציוני המרצה ייתכן ומשקף הבנה אנושית משותפת של אמות מידה איכותיות, גם אם קיימים הבדלים ברמת המומחיות. הערכת העמיתים נמצאה קרובה יותר לזו של המרצה, ממצא הנתמך על ידי מחקרים קודמים שהצביעו על הערכות עמיתים כבאלו העשויות להשתוות במצבים מסוימים להערכות מומחה (Double et al., 2020; Li et al., 2019; Usher, 2025).

ממצא מרכזי נוסף קשור להבחנה בין רמות שונות של איכות הפרויקט המוערך. נמצא כי ככל שאיכות הפרויקט עלתה, כך השתפרה ההלימה בין ציוני ChatGPT לציוני המרצה – כאשר הפער הגדול ביותר נרשם עבור פרויקטים חלשים. למעשה, לא נמצא קשר מובהק בין ציוני ChatGPT לציוני המרצה עבור פרויקטים ברמה נמוכה או בינונית, בעוד שנמצא קשר חיובי יחסית חזק עבור הפרויקטים שהיו מלכתחילה באיכות גבוהה. דפוס זה משקף ממצאים ממחקר קודם על פיו ביצועי GenAI הם הנמוכים ביותר בעת הערכת תוצרים של קבוצות חלשות (Gorgun & Yildirim-Erbasli, 2024; Gurin et al., 2025). ייתכן שפער זה משקף קושי לזהות ליקויים מהותיים ונקודות לשיפור בעבודות חלשות, או מרמז שאיתורם אינו בהכרח מתורגם לכדי הערכה מדויקת. לעומת זאת, ההלימה הבינונית עם ציוני המרצה שנמצאה בעבודות באיכות הגבוהה ביותר עשויה להעיד על יעילות גבוהה יותר של ChatGPT בעת הערכת עבודות שהן מלכתחילה מלוטשות, מאורגנות וקוהרנטיות – מאפיינים המזוהים לעיתים קרובות עם עבודה אקדמית איכותית.

בחינת הערכות העמיתים גם כן הציגה דפוס של תלות באיכות הפרויקט המוערך: פרויקטים חלשים יותר זכו לרוב לציונים גבוהים יתר על המידה, בעוד שפרויקטים חזקים הוערכו באופן שמרני ולעיתים בחומרה יתרה. תופעה זו עולה בקנה אחד עם עדויות המצביעות על קושי של סטודנטים לבקר באופן ביקורתי עבודות איכותיות – בין אם בשל חוסר ביטחון, חוסר הכשרה, או סולידריות עם חבריהם לביתה (Suñol et al. 2016; Topping, 2005; Usher & Barak, 2018). מחקר קודם מצא באופן דומה כי סטודנטים הציעו משוב פחות איכותי ככל שאיכות העבודה של עמיתיהם הלכה ועלתה, ייתכן בשל תחושת אי-מחויבות לתמוך (או נוחות רבה לבקר) עמיתים בעלי ביצועים גבוהים (Banihashem et al., 2024). גורמים קוגניטיביים ייתכן וגם הם שיחקו תפקיד פה: עבודה באיכות נמוכה יותר מציגה שגיאות גלויות יותר, בעוד שהערכת פרויקטים חזקים דורשת ידע דיסציפלינרי רב ושיקול דעת מעמיק – אתגר שעלול להקשות על סטודנטים חסרי ניסיון משמעותי בהערכה.

לסיכום, ממצאי המחקר מדגישים את הסיכון הפוטנציאלי שבהסתמכות יתר על הערכה מבוססת GenAI ובפרט כאשר זו משמשת כמערכת העומדת בפני עצמה. הערכה זו עלולה להטעות או לבלבל את הסטודנטים, במיוחד אם זו לא תלווה בבחינה ביקורתית ומושכלת של תוכנה וטיבה. נטיית ChatGPT להערכת־יתר, במיוחד עבור עבודות באיכות נמוכה, עלולה לפגוע בתוקף התהליך ההערכתי – ומכאן הצורך לנקוט במשנה זהירות בעת שילוב מערכות אלו כחלק מהערכה מסכמת או כהערכה בלעדית במסגרת קורסים אקדמיים. עם זאת, כלים דוגמת ChatGPT עשויים לתרום רבות לתהליכי הערכה מעצבת (הערכה לשם למידה), בהם משוב מפורט וממוקד עשוי לתמוך בשיפור הדרגתי. דפוסי המהימנות התלויים באיכות העבודה המוערכת שדווחו במחקר זה מדגישים את הסיכון שבהחלת מודל הערכה אחיד וסטנדרטי על כלל הסטודנטים: הערכת־יתר של עבודות חלשות עלולה ליצור אשליית הצלחה ולפגוע בתהליכי שיפור, בעוד שהערכת־חסר של עבודות חזקות עלולה לפגוע במוטיבציה ובביטחון של סטודנטים מצטיינים.

ממצאים אלו תומכים בפיתוח אסטרטגיות הערכה אדפטיביות, המבוססות על שילוב בין החוזקות הייחודיות של כל אחד ממקורות ההערכה: GenAI והערכת עמיתים. כך למשל, הערכת עמיתים עשויה לסייע בזיהוי ליקויים בסיסיים בשלבי עבודה מוקדמים, בעוד ש- GenAI עשוי לספק משוב טכני ומובנה בשלבים מתקדמים יותר של העבודה. חשיפת סטודנטים למקורות הערכה מגוונים יכולה לעודד חשיבה ביקורתית על אמות מידה איכותיות, ולהגביר את מעורבותם בתהליך הלמידה – בייחוד כאשר נעשה שימוש פדגוגי ביכולת הדיאלוגית של GenAI לצורך יצירת אינטראקציות לימודיות מותאמות אישית. יישום יעיל של מודלים מעין אלו יצריך הכשרה מוקדמת וממוקדת לסטודנטים – הן בבלי הערכת עמיתים והן בפיתוח מיומנויות של הנדסת פרומפטים ואוריינות AI, לשם מיצוי הפוטנציאל הקיים בשילוב GenAI בהקשרי הערכה לימודית. ברמה המוסדית, ניתן לפתח מודלים רציפים של הערכה אינטגרטיבית, שבהם GenAI משמש להערכת דיוק טכני, הערכת עמיתים מספקת תובנות הקשריות, והמרצה מספק את הסינתזה המקצועית הסופית. מודלים מעין אלו עשויים לקדם תהליך הערכה שוויוני, שקוף ומשמעותי יותר מבחינה פדגוגית.

מקורות

Banihashem, S.K., Kerman, N.T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education 21*(23). https://doi.org/10.1186/s41239-024-00455-4

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education, 20*(1), 43 .https://doi.org/10.1186/s41239-023-00411-8

Double, K.S., McGrane, J.A., & Hopfenbeck, T.N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review 32*(2), 481-509. https://doi.org/10.1007/s10648-019-09510-3

Essel, H. B., Vlachopoulos, D., Tachie-Menson, A., Johnson, E. E., & Baah, P. K. (2022). The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education. International *Journal of Educational Technology in Higher Education, 19*(57). https://doi.org/10.1186/s41239-022-00362-6

Falchikov, N., & J. Goldfinch. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research 70*(3), 287-322.

Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal 51*(1), 201-224. https://doi.org/10.1002/berj.4069

Gorgun, G. & Yildirim-Erbasli, S.N. (2024). Algorithmic bias in BERT for response accuracy prediction: A case study for investigating population validity. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12420

Gurin S., Klebanov, B., & Alexandron, G. (2025). Uncovering measurement biases in LLM Embedding spaces: The Anna Karenina Principle and its implications for automated feedback. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-025-00485-7

Haudek, K.C. & Zhai, X. (2023). Examining the effect of assessment construct characteristics on machine learning scoring of scientific argumentation. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00385-8

Labadze, L., Grigolia, M. & Machaidze, L. (2023). Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education, 20*(56). https://doi.org/10.1186/s41239-023-00426-1

Li, H., Xiong, Y., Hunter, C.V., Guo, X., & Tywoniw, R. (2020) Does peer assessment promote student learning? A meta-analysis, *Assessment & Evaluation in Higher Education, 45*(2), 193-211. https://doi.org/10.1080/02602938.2019.1620679

Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024) Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment & Evaluation in Higher Education, 49*(5), 616-633. DOI: 10.1080/02602938.2024.2301722

Morris, W., Holmes, L., Choi, J.S. & Crossley, S. (2024). Automated Scoring of Constructed Response Items in Math Assessment Using Large Language Models. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00418-w

Ocampo, J.C., Panadero, E., Zamorano, D., Sánchez-Iglesias, I., & Ruiz, F.D. (2024) The effects of gender and training on peer feedback characteristics. *Assessment & Evaluation in Higher Education, 49*(4), 539-555, DOI: 10.1080/02602938.2023.2286432

Okonkwo, C. W., & Ade-Ibijola, A. O. (2021). Chatbots applications in education: A systematic review. *Computers & Education: Artificial Intelligence, 2*, 100033. https://doi.org/10.1016/j.caeai.2021.100033

Pinto, G., Cardoso-Pereira, I., Monteiro, D., Lucena, D., Souza, A., & Gama, K. (2023). Large language models for education: Grading open-ended questions using ChatGPT. In *37th Brazilian Symposium on Software Engineering*, edited by Edna Dias Canedo, 293-302. New York, NY, USA: ACM. https://doi.org/10.1145/3613372.3614197

Suñol, J. J., Arbat, G., Pujol, J., Feliu, L., Fraguell, R.M., & Planas-Lladó, A. (2016). Peer and self-assessment applied to oral presentations from a multidisciplinary perspective. *Assessment & Evaluation in Higher Education, 41*(4), 622-637. https://doi.org/10.1080/02602938.2015.1037720

Tam, A. C. F. (2024). Interacting with ChatGPT for internal feedback and factors affecting feedback quality. *Assessment & Evaluation in Higher Education*, 1-17. https://doi.org/10.1080/02602938.2024.2374485

Topping, K.J. (2005). Trends in peer learning. *Educational Psychology 25*(6), 631-645. https://doi.org/10.1080/01443410500345172

Topping, K.J., Gehringer, E., Khosravi, H., Gudipati, S., Jadhav, K., & Susarla, S. (2025). Enhancing peer assessment with artificial intelligence. *International Journal of Educational Technology in Higher Education 22*(3). https://doi.org/10.1186/s41239-024-00501-1

Usher, M., & Barak, M. (2018). Peer Assessment in a Project-Based Engineering Course: Comparing between on-Campus and Online Learning Environments. *Assessment & Evaluation in Higher Education 43*(5), 745-759. https://doi.org/10.1080/02602938.2017.1405238.

Usher, M. (2025). Generative AI vs. Instructor vs. Peer Assessments: A Comparison of Grading and Feedback in Higher Education. *Assessment & Evaluation in Higher Education 50*(6), 912-927. https://doi.org/10.1080/02602938.2025.2487495.

Usher, M., Barak, M., & Erduran, S. (2025). What Role Should Higher Education Institutions Play in Fostering AI Ethics? Insights from Science and Engineering Graduate Students. *International Journal of STEM Education 12*(1). https://doi.org/10.1186/s40594-025-00567-x.

Venter, J., Coetzee, S.A., & Schmulian, A. (2024). Exploring the use of artificial intelligence (AI) in the delivery of effective feedback. *Assessment & Evaluation in Higher Education 50* (4): 516-536. https://doi.org/10.1080/02602938.2024.2415649