צ׳טבוטים מבוססי AI מול מעריכים אנושיים: ניתוח דיוק ציונים ואיכות משובים בהשכלה הגבוהה

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

AI Chatbots vs. Human Assessors: A Comparative Analysis of Grading Accuracy and Feedback Quality in Higher Education

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

Abstract

In recent years, the integration of Generative Artificial Intelligence (GenAI) in education has introduced innovative approaches to assessment. One such approach is AI chatbot-based assessment, which utilizes large language models (LLMs) to provide students with timely and consistent feedback. However, the effectiveness of AI chatbots in generating assessments comparable to human evaluators remains unclear, with limited research offering direct comparisons in educational contexts. This study aims to evaluate and compare the grading accuracy and feedback quality of AI chatbots-based assessments with those provided by a course instructor and peers in a higher education setting. The participants were 76 undergraduate students who engaged in a group project involving three phases: questionnaire development, peer assessment, and chatbot-based assessment. Employing a mixed-methods approach, data were collected through a quantitative comparison of project grades and a qualitative analysis of feedback types and quality. Results indicated that AI chatbots consistently assigned higher grades than human assessors, while peer and instructor grades were similarly lower and closely aligned. Content analysis revealed that AI chatbots generally provided higher-quality feedback compared to peers, often offering detailed insights and specific guidance for improvement. However, there were instances where AI chatbot feedback was irrelevant or contradictory to course content. In contrast, peer feedback tended to be more personalized and context-specific. These findings highlight the importance of human judgment, suggesting that integrating chatbot-based assessments with traditional methods could leverage their complementary strengths to enhance student learning.

Keyword: Chatbot-based assessment, Generative Artificial Intelligence (GenAI), Higher education, Peer assessment, Peer feedback.

תקציר

בשנים האחרונות, שילוב בינה מלאכותית ג׳נרטיבית (GenAI) בחינוך פותח אפשרויות חדשות ללמידה ולהערכה. אחת מהאפשרויות הללו היא הערכה מבוססת צ׳טבוטים, המתייחסת לשימוש במודלי שפה גדולים (LLMs) בכדי לספק לסטודנטים משוב עקבי בזמן אמת. עם זאת, יעילותם של צ׳טבוטים בהערכת תוצרי למידה של סטודנטים טרם נבחנה דייה וקיים מחקר מוגבל המציע השוואות ישירות בין הערכה זו לבין הערכה אנושית, בעיקר בהקשרים חינוכיים.

מטרת המחקר הנוכחי הינה לבחון את יעילותם של צ׳טבוטים בהערכת פרויקטים קבוצתיים של סטודנטים, בהשוואה להערכות המרצה והערכות העמיתים בקורס בהשכלה הגבוהה. יעילות זו נסקרה הן מבחינת מידת דיוק הציונים והן מבחינת איכות המשובים שניתנו לפרויקטים, בהתבסס על אותו המחוון. במחקר השתתפו 76 סטודנטים לתואר ראשון, אשר לקחו חלק בפרויקט קבוצתי שכלל שלושה שלבים: פיתוח שאלון, הערכת עמיתים והערכה מבוססת צ׳טבוט. הנתונים נאספו דרך השוואה כמותית של ציוני הפרויקטים וניתוח איכותני של סוגי המשובים ואיכותם. הממצאים הראו כי צ׳טבוטים העניקו ציונים ממוצעים גבוהים יותר מן המעריכים האנושיים, בעוד שציוני המרצה והעמיתים היו נמוכים יותר, אך בקורלציה גבוהה האחד עם השני. ניתוח התוכן הראה כי צ׳טבוטים סיפקו לרוב משוב באיכות גבוהה יותר בהשוואה לעמיתים, ונטו להציע תובנות מפורטות לצד הנחיות ספציפיות וישימות לשיפור הפרויקטים. עם זאת, היו מקרים בהם המשוב שהופק על ידי צ׳טבוטים נמצא כלא רלוונטי או כסותר את התוכן הנלמד בקורס. מאידך, משוב העמיתים נטה להיות יותר ממוקד בהקשרו מבחינת התאמה לתכני הקורס. תוצאות אלו מציעות כי שילוב צ׳טבוטים לצד שיטות הערכה מסורתיות יותר עשוי לנצל את החוזקות המשלימות של כל מתודה, תוך הדגשת חשיבות השיפוט האנושי בתהליכי הערכה בחינוך.

מילות מפתח: בינה מלאכותית ג׳נרטיבית (GenAI), הערכה מבוססת צ׳טבוט, הערכת עמיתים, השכלה גבוהה, משובי עמיתים.

מבוא

שילובה ההולך וגובר של בינה מלאכותית ג׳נרטיבית (GenAI) במסגרות חינוכיות פתחה אפיקים חדשים לשיטות הערכה חדשניות (Usher & Barak, 2024; Chan & Hu, 2023; Tam, 2024). בין היישומים המבטיחים ביותר של GenAI בהקשר להערכה הוא השימוש בצ׳טבוטים – מתווכים אוטומטיים הפועלים באמצעות מודלי שפה גדולים 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).

מחקרים הדגישו את הפוטנציאל, כמו גם את האתגרים, הכרוכים בשילוב צ׳טבוטים מבוססי AI בתהליכי הערכה בהשכלה הגבוהה (Farazouli et al., 2023). צ׳טבוטים אלו מתוארים כמעין עוזרי הוראה וירטואליים העשויים לסייע לסטודנטים במשימות לימודיות שונות, לספק תשובות לשאלותיהם ולהבהיר מושגים מורכבים (Ding et al., 2023; Labadze et al., 2023). הם נתפסים ככלי המספק משוב אינטראקטיבי ומיידי, סיוע מותאם אישית ותמיכה מקיפה (Chan & Hu, 2023; Okonkwo & Ade-Ibijola, 2021). בנוסף, צ׳טבוטים יכולים לשמש כקיצור דרך מועיל, ולהפחית באופן משמעותי את הזמן הנדרש להשלמת משימות אקדמיות (Chan & Hu, 2023; Labadze et al., 2023). מעבר לסיוע לסטודנטים, צ׳טבוטים עשויים לתמוך גם בצוותי ההוראה על ידי בדיקת מטלות, הפקת מערכי שיעור וחומרי למידה אינטראקטיביים (Walter, 2024). שילוב זה מקל על העומס הרב המוטל על מורים ומרצים ובכך מאפשר להם להתמקד במשימות הוראה מורכבות יותר (Ding et al., 2023; Essel et al., 2022).

למרות יתרונות אלו, קיים הסיכון שצ׳טבוטים יפרשו באופן חלקי או שגוי את תוצרי הסטודנטים, מה שעלול להוביל למשוב שאינו רלוונטי או שאינו מדויק (Chan & Hu, 2023). בנוסף, הסטודנטים חייבים להיות בעלי רמת אוריינות מסוימת ב- AI בכדי לעשות שימוש יעיל ומושכל בכלים אלו, שכן איכות המשוב שנוצר תלוי במידה רבה במידת הבהירות והדיוק של ההנחיות הניתנות (Tam, 2024; Walter, 2024). לבסוף, קיימת אי-וודאות בנוגע למידת הקורלציה בין הערכות שנוצרו על ידי צ׳טבוטים לבין הערכות אנושיות (Lu et al., 2024). מחקרים הראו כי מודלים ממוחשבים מבוססי AI מסוגלים לחזות ציונים ברמת דיוק המתקרבת לזו של מומחים (Haudek & Zhai, 2023; Morris et al., 2024). לדוגמה, מחקר שנערך לאחרונה מצא עקביות בינונית עד טובה ביעילותם של צ׳טבוטים במתן ציון לעבודות סטודנטים, בהשוואה להערכות המרצים (Lu et al., 2024). עם זאת, השוואות בין הערכות הניתנות על ידי AI להערכות אנושיות ממשיכות לעורר חששות משמעותיים בכל הנוגע לדיוק ולעקביות (Okonkwo & Ade-Ibijola, 2021).

מחקרים קודמים בחנו גם את הקשרים בין הערכות הניתנות על ידי מקורות אנושיים שונים, כגון הערכות עמיתים מול הערכות מרצים. מחקרים דיווחו כי הערכות עמיתים יכולות להיות אמינות כמו הערכות מרצים, בעיקר כשמספקים מחוון ברור ומציעים לסטודנטים הכשרה מתאימה (Usher & Barak, 2018; Li et al., 2020). עם זאת, ישנם מקרים בהם הערכות עמיתים עשויות להיות שונות מהערכות המרצים, במיוחד מבחינת עקביות ועומק המשובים הכתובים (Falchikov & Goldfinch, 2000; Suñol et al., 2016). מצב

זה מעלה שאלות בנוגע למידת ההתאמה בין סוגי הערכות שונים – בין אם מבוססות צ׳טבוטים, הערכות מרצים או עמיתים (Lu et al., 2024; Tam, 2024).

מטרה ושאלות המחקר

מטרת המחקר הינה לבחון ולהשוות את מידת דיוק הציונים ואיכות המשובים הניתנים לפרויקטים קבוצתיים של סטודנטים משלושה מקורות: צ׳טבוטים מבוססי AI, מרצת הקורס והעמיתים.

מכאן עלו שאלות המחקר הבאות:

1. מהי מידת דיוק הציונים המופקים על ידי צ׳טבוטים אל מול ציוני המרצה והעמיתים עבור פרויקטים קבוצתיים של סטודנטים?
2. כיצד סוגי ואיכות המשובים המופקים על ידי צ׳טבוטים שונים מאלו הנכתבים על ידי העמיתים עבור פרויקטים קבוצתיים של סטודנטים?

אוכלוסיית וסביבת המחקר

במחקר השתתפו 76 סטודנטים לתואר ראשון (89% נשים, 11% גברים) שהשתתפו בקורס העוסק במדידה והערכה בשנת הלימודים 2023-24. כחלק מדרישות הקורס, הסטודנטים עבדו על פרויקט בקבוצות של 3-4 משתתפים, אשר חולק לשלושה שלבים: פיתוח שאלון, הערכת עמיתים, והערכה מבוססת צ׳טבוט. ראשית, הסטודנטים עבדו בקבוצות על פיתוח שאלון מקוון בנושא לבחירתם, שכלל לפחות ארבע שאלות סגורות ולפחות שתי שאלות פתוחות. בשלב השני, כל סטודנט ביצע הערכת עמיתים לשני פרויקטים של חבריו לכיתה. התהליך היה אנונימי, והסטודנטים השתמשו במחוון מפורט שחולק לשישה חלקים: מטרת המחקר, אוכלוסיית המחקר, שאלות המחקר, שאלות סגורות בשאלון, שאלות פתוחות בשאלון והקדמה לשאלון. עבור כל אחד משש חלקי המחוון צוינו הנחיות מילוליות וכן ציון כמותי מתוך הציון הכולל. הסטודנטים העניקו ציונים מספריים ומשוב כתוב עבור כל אחד משש החלקים. לבסוף, הסטודנטים ביצעו הערכה מבוססת צ׳טבוטים, בה התבקשו לבחור צ׳טבוט לפי העדפתם (דוגמת ChatGPT או Gemini) ולהפיק באמצעותו הערכה שהתבססה על אותו המחוון בכדי לספק ציונים מספריים ומשוב כתוב. מרצת הקורס העריכה את השאלונים כשלב אחרון, לאחר שהוגשו כל חלקי הפרויקט.

שיטת המחקר, כלי המחקר וניתוח

במחקר נעשה שימוש במתודת המחקר המשולב המקביל. הרכיב הכמותי כלל את השוואת הציונים שניתנו לפרויקטים על ידי צ׳טבוטים, המרצה והעמיתים, תוך שימוש באותו המחוון בעל ששת הקריטריונים. הנתונים נותחו באמצעות סטטיסטיקה תיאורית ומבחן ANOVA חד-כיווני למדידות חוזרות על מנת להעריך הבדלים בממוצעי הציונים בין שלוש קבוצות המעריכים. בנוסף, נערכו מבחני t למדגמים מזווגים בכדי להשוות את ממוצעי הציונים שניתנו על ידי כל זוג מעריכים. לבסוף, בוצעו מבחני מתאם מסוג פירסון על מנת להעריך את הקשרים בין ציוני הפרויקטים ובמטרה לקבוע את עוצמת וכיוון הקשרים בין מקורות ההערכה השונים.

הרכיב האיכותני כלל ניתוח תוכן של סוגי ואיכות המשובים הכתובים משני מקורות: הערכות עמיתים והערכות הצ׳טבוטים. הסטודנטים רשמו משוב נפרד עבור כל אחד משש הקריטריונים במחוון, עבור כל אחד משני הפרויקטים שהעריכו. הצ׳טבוטים יצרו משוב עבור כל הפרויקט, גם כן בהתאם לאותם הקריטריונים. תהליך זה הניב סך של 1,368 הערות משוב – 912 מהערכות העמיתים ו-456 מהערכות הצ׳טבוטים. המשובים דורגו לפי רמת איכותם על סולם הנע בין 1 (נמוך ביותר) ל-5 (גבוה ביותר), בהתבסס על סטנדרטים שנקבעו במחקר קודם (Usher & Barak, 2018). הסולם מפורט כדלהלן: חיזוק (1 נקודה) – הערות כלליות ללא כל תרומה קוגניטיבית לשיפור הפרויקט המוערך. הצהרה (2 נקודות) – הערות המציינות עמידה או אי עמידה בהנחיות הפרויקט על פי המחוון, ללא ניתוח מעמיק או הצעות לשיפור. אימות (3 נקודות) – הערות המצביעות על חלקים מסוימים טעוני שיפור בפרויקט המוערך, אך ללא הסבר מפורט ו/או ניתוח מעמיק. הרחבה (4 נקודות) – הערות אינפורמטיביות המספקות תובנות מפורטות והסבר אילו חלקים יש לשפר ומדוע השיפורים נחוצים. הכוונה (5 נקודות) – הערות באיכות גבוהה המספקות תובנות מפורטות והסבר אילו חלקים יש לשפר, מדוע השיפורים נחוצים ואף הצעות ספציפיות ליישום השיפורים המוצעים. הניתוח האיכותני נעשה על ידי מרצת הקורס וכותבת המאמר. לבחינת מהימנות תהליך הניתוח, נבחרה באקראי דגימה של כ-10% מהמשובים והוערכה על ידי שתי שופטות נוספות, סטודנטיות לדוקטורט

בחינוך מדעי. נמצא מתאם של כ- 90% הסכמה בשני המקרים. להשוואת התפלגות רמות איכות המשוב בין העמיתים לבין הצ׳טבוטים, בוצעה סדרת מבחני חי בריבוע.

ממצאים

הבדלים במידת דיוק הציונים שניתנו לפרויקטים על ידי צ׳טבוטים, המרצה והעמיתים

הציונים הממוצעים שהופקו על ידי הצ׳אטבוטים היו באופן עקבי גבוהים יותר ($M = 92.04$, $SD = 6.74$) בהשוואה לציוני מרצת הקורס והעמיתים שהיו נמוכים יותר אך קרובים ($M = 84.24$, $SD = 6.78$; $M = 83.47$, $SD = 8.45$). ניתוח ANOVA הצביע על הבדל מובהק סטטיסטית בין ציוני שלוש הקבוצות ($F(1.58, 123.10) = 52.95$, $p < .001$; $\eta^2 = .40$). בסדרת מבחני t נמצאו הבדלים מובהקים בין הציונים שהוקצו על ידי צ׳אטבוטים אל מול ציוני העמיתים ($t(75) = -7.52$, $p < .001$, $d = .85$) ומול ציוני המרצה ($t(75) = -9.44$, $p < .001$, $d = 1.06$). עם זאת, ההבדל בין הציונים שניתנו על ידי העמיתים לבין ציוני המרצה לא היה מובהק סטטיסטית ($t(75) = .86$, $p = .39$, $d = .09$). מבחן פירסון חשף מספר קשרים מרכזיים בין שלושת מקורות ההערכה. כפי שניתן לראות בטבלה 1, נמצא מתאם חיובי חזק ומובהק בין הציונים שניתנו על ידי העמיתים לבין אלו שניתנו על ידי המרצה ($r(74) = .63$, $p < .001$). בין ציוני הצ׳אטבוטים לציוני המרצה נמצא מתאם חיובי בינוני ($r(74) = .42$, $p < .001$), בעוד שבין ציוני הצ׳אטבוטים לעמיתים נמצא מתאם חלש ולא מובהק ($r(74) = .14$, $p = .22$).

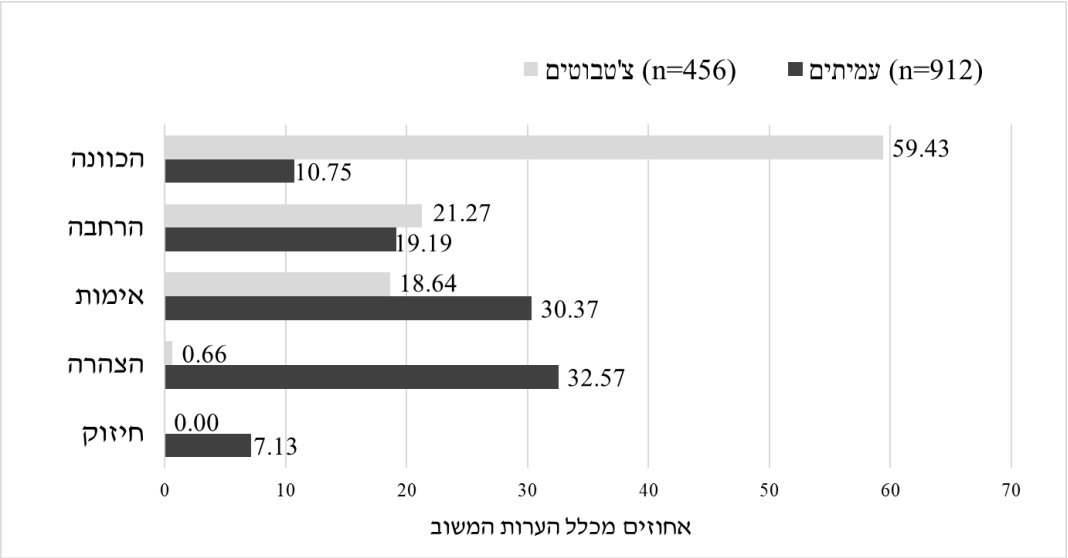
טבלה 1. מתאמים עבור ציוני הצ׳טבוטים, המרצה והעמיתים

מקור ההערכה	צ׳טבוטים	מרצה	עמיתים
צ׳טבוטים	1	-	-
מרצה	0.42***	1	-
עמיתים	0.14	0.63***	1

הבדלים באיכות המשובים שניתנו לפרויקטים על ידי צ׳טבוטים מול העמיתים

הניתוח האיכותני העלה כי הצ׳טבוטים סיפקו בממוצע משובים באיכות גבוהה יותר בהשוואה לעמיתים. איור 1 מציג את התפלגות הערות המשוב שסופקו על ידי הצ׳אטבוטים והעמיתים לפי חמש רמות האיכות שנבחנו, באחוזים מתוך כלל הערות המשוב שסופקו על ידי כל אחד ממקורות ההערכה הללו.

נמצא כי משובים מסוג *חיזוק*, שהוגדרו כמשוב באיכות הנמוכה ביותר, הופיעו ב- 7.13% בלבד מהערות המשוב שניתנו על ידי העמיתים ונעדרו לחלוטין ממשוב הצ׳טבוטים. דוגמאות להערות מסוג זה כללו הכרה כללית בחוזקות הפרויקט המוערך: ״מבוא מצוין״ (S1, גבר) או ״עבודה ממש טובה, אין לי מה להוסיף!״ (S37, אישה). באופן דומה, משובים מסוג *הצהרה* היו גם כן נפוצים בעיקר בקרב העמיתים (32.57%), בעוד שפחות מ- 1% מהם נוצרו על ידי צ׳טבוטים, עם הבדל מובהק סטטיסטית ($\chi^2 = 119.29$, $p < .001$). הערות אלו כללו הצהרה מה קיים או חסר בפרויקט לפי הקריטריונים במחוון, מבלי לספק הסבר נוסף, לדוגמה: ״אין התייחסות לסולם המדידה של כל שאלה כפי שנדרש בקריטריונים להערכה״ (S63, אישה).



איור 1. התפלגות המשובים לפי חמש רמות האיכות (%), עבור צ׳טבוטים ועמיתים

משובים מסוג *אימות*, שנחשבים למשוב באיכות בינונית, היו כמעט כפולים בשכיחותם בקרב העמיתים (30.37%) בהשוואה לצ׳טבוטים (18.64%), ($\chi^2 = 62.08$, $p < .001$). למרות שההערות אלו לרוב זיהו מקומות ספציפיים בפרויקט שניתן לשפרם, הן חסרו הסברים מפורטים מדוע השיפורים נחוצים או כיצד ליישמם. כך למשל, סטודנטית הציעה כי ״יש להוסיף שאלות פתוחות נוספות המתייחסות להיבטים ספציפיים יותר של השימוש בכלי בינה מלאכותית בקורסי תכנות״ (S16, אישה). בנוסף, משובי העמיתים כללו לעיתים קרובות הצעות מעורפלות או מהוססות, שקדמו להן ביטויים דוגמת ״לדעתי״ או ״אולי״, כפי שניסחה זאת סטודנטית אחת: ״אוכלוסיית המחקר שבחרת די רחבה; אולי כדאי לדייק קצת יותר״ (S31, אישה). צ׳טבוטים גם כן סיפקו הערות מסוג זה, עם זאת, לעיתים אלו היו כלליות מדי וחסרו קשר ברור לפרויקט המוערך. לדוגמה, בשיחה אחרת, הופק על ידי צ׳טבוט המשוב הבא: ״ניסחת שאלות סגורות באיכות גבוהה [..] עם זאת, ניתן לשפר מעט את השאלון על ידי שיפור השאלות הסגורות [..] להלן טיפים לכתיבת שאלות סגורות: וודא שהשאלות מנוסחות בצורה ברורה, תמציתית ולא מוטה; השתמש בשפה פשוטה וקלה להבנה; נסה לנסח שאלות בצורה ספציפית ככל האפשר [..]״ (הופק על ידי ChatGPT).

מנגד, משובים מסוג *הרחבה*, שהוגדרו כמשוב באיכות בינונית-גבוהה, היו פחות שכיחים בקרב העמיתים (19.19%) בהשוואה לצ׳טבוטים (21.27%), ($\chi^2 = 27.01$, $p < .001$). משובים אלו סיפקו לרוב תובנות מפורטות הכוללות הסברים מדוע נחוץ לשפר חלקים בפרויקט. דפוס בולט במשובי העמיתים הייתה הנטייה לחזור בהערכותיהם על תוכן הדיונים שנערכו בכיתה: ״השאלה הפתוחה שלך [..] רחבה מדי. היא מתייחסת גם לשינויים שהמשתתפים חווים וגם לשינויים שהם היו רוצים לחוות, דבר שעלול לבלבל את הנבדקים ולהוביל לתשובות לא ממוקדות. יש כאן שתי שאלות באחת, והמרצה דיברה במפורש על כך שיש להימנע מכך בשיעור״ (S42, אישה). המשובים שהופקו על ידי הצ׳טבוטים תחת קטגוריית ההרחבה היו לעיתים מפורטים יותר מאלו של העמיתים. עם זאת, היו מקרים בהם צ׳אטבוטים הציעו הצעות מפורטות שלא היו מדויקות לחלוטין או שלא תאמו את הנלמד בכיתה, מה שהוביל לעיתים את הסטודנטים לתקן את הצ׳טבוט. במקרה אחד, הופקה על ידי צ׳אטבוט ההערה הבאה: ״שאלות המחקר מנוסחות בבירור [..] עם זאת, ניתן לשפר מעט את החלק על ידי דיוק נוסף של שאלות המחקר, הוספת שאלות נוספות, ופירוט נוסף של משתני המחקר. דוגמה לשאלת מחקר משופרת: ״איזו רמת רעש מעדיפים סטודנטים לתואר ראשון בזמן הלימודים?״ (הופק על ידי Gemini). הסטודנטית הגיבה כך: ״הדוגמה שלך לשאלת מחקר משופרת לא מדויקת. אנחנו צריכים לנסח שאלת מחקר כיחסים בין משתנים. ההצעה שלך אינה תואמת את ההנחיות שקיבלנו בכיתה וגם אינה רלוונטית למטרת המחקר שהצגנו״ (S4, אישה).

לבסוף, משובים מסוג *הכוונה*, שהוגדרו כמשוב באיכות הגבוהה ביותר, היו נפוצים כמעט פי שישה בקרב צ׳אטבוטים (59.43%) מאשר בקרב העמיתים (10.75%), ($\chi^2 = 92.55$, $p < .001$). משובי הכוונה שנכתבו על ידי עמיתים, כאשר היו נוכחים, היו מותאמים אישית ובעלי הקשר, במיוחד בהתייחס לרלוונטיות הפרויקט לקורס, לדוגמה: ״ניסוח מטרת המחקר בעייתי בשל השימוש במילה ׳השפעה׳. לא ניתן לבדוק את ההשפעה בין משתנים בסוג זה של מחקר [..]. המרצה אפילו הדגישה את זה כמה פעמים בשיעור. אני מציעה לנסח מחדש כדי לבחון את הקשר בין המשתנים, משהו כמו ׳מהו הקשר בין מרחבי למידה שונים לבין רמת הריכוז

של סטודנטים׳׳ (S69, אישה). משובים שהופקו על ידי צ׳טבוטים וקוטלגו כהכוונה הציגו תובנות מפורטות לצד הנחיות ספציפיות לביצוע השיפורים המוצעים, למשל: ״ניסוח האוכלוסייה אינו מבהיר בבירור אילו מורים ייכללו במחקר [..] המלצות: ציין בבירור אילו מורים ייכללו [..]. אוכלוסיית המחקר יכולה להיות מנוסחת כך: ׳אוכלוסיית המחקר כוללת מורים המועסקים במשרה מלאה או חלקית בבתי ספר יסודיים וחטיבות ביניים ברחבי הארץ. המחקר יתמקד במורים מכל תחומי הדעת והוותק בהוראה. ייעשה מאמץ להבטיח ייצוג הולם של מורים מכל המגזרים והרמות, תוך התחשבות במגדר, גיל וותק׳׳ (הופק על ידי Gemini).

עם זאת, היו מקרים בהם משוב שהופק על ידי צ׳טבוטים נמצא כלא רלוונטי לפרויקט המוערך ולעתים אף סתר מידע שנלמד במהלך ההרצאות. כך לדוגמה, במקרה הבא, הצ׳טבוט הפיק משוב ובו הציע לסטודנט להוסיף השערת מחקר, דבר שלא נדון בכיתה: ״מטרת המחקר ברורה וניתנת להבנה בקלות [..] עם זאת, ניתן לשפר את ניסוח המטרה על ידי הצגת השערה ספציפית. השערה מוגדרת היטב לא רק ממקדת את המחקר, אלא גם מספקת כיוון ברור לאיסוף וניתוח הנתונים [..] לדוגמה, ניתן להעלות השערה כי אנשים שמחזיקים חיות מחמד נוטים יותר לאינטראקציות חברתיות פנים אל פנים מאשר אלו שלא מחזיקים חיות מחמד״ (הופק על ידי ChatGPT). במספר מקרים אחרים, צ׳טבוטים הציעו לפרט את אופי הקשר בשאלת המחקר, דבר שסותר את העקרונות שנלמדו במהלך הקורס, לדוגמה: ״ניתן לשפר את ניסוח שאלת המחקר על ידי הצגת אופי הקשר בין המשתנים במפורש. ניסוח מוצע: האם קיים קשר חיובי בין בעלות על כלב לבין תדירות ומגוון האינטראקציות החברתיות של בעליו?״ (הופק על ידי Gemini). סוג זה של המלצות לא רלוונטיות או מנוגדות לנלמד בקורס היה בולט בהיעדרו ממשובי העמיתים.

דיון

ממצאי המחקר מספקים תובנות לגבי מידת הדיוק ואיכות ההערכות הניתנות לפרויקטים קבוצתיים של סטודנטים מצד שלושה מקורות: צ׳טבוטים מבוססי בינה מלאכותית, מרצת הקורס והעמיתים. ממצא מרכזי הצביע על פער במידת דיוק הציונים שניתנו לאותם פרויקטים מצד שלושת סוגי המעריכים. התוצאות לימדו כי הצ׳טבוטים נקטו בגישה פחות מחמירה לציינון הפרויקטים, בעוד שהמעריכים האנושיים – בין אם המרצה או העמיתים – נטו ליישם קריטריונים מחמירים יותר. בהמשך לכך, נמצא מתאם חיובי חזק ומובהק בין ציוני המרצה והעמיתים, המעיד על רמת הסכמה גבוהה בין שתי קבוצות אלו. ממצא זה תואם מחקרים קודמים המראים כי בהקשרים מסוימים הערכות עמיתים יכולות להיות מדויקות כמו הערכות מרצים, במיוחד כאשר ישנם קריטריונים ברורים להערכה והכשרה מתאימה (Usher & Barak, 2018; Li et al., 2020; Ocampo et al., 2024). לעומת זאת, נמצא מתאם נמוך ולא מובהק בין ציוני העמיתים והצ׳טבוטים, ממצא העולה בקנה אחד עם החששות לגבי מידת הדיוק והעקביות של הערכות שנוצרו על ידי בינה מלאכותית, במיוחד כשמדובר בהערכת תוצרים מורכבים ואיכותיים (Lu et al., 2024; Okonkwo & Ade-Ibijola, 2021).

ממצא מרכזי נוסף קשור לאיכות המשובים שהופקו על ידי הצ׳טבוטים בהשוואה לעמיתים. ניתוח התוכן חשף הבדלים מובהקים בין שני המקורות הללו, כאשר הצ׳טבוטים נטו לספק משובים באיכות גבוהה יותר בהשוואה לעמיתים. מעל ל- 60% מהמשובים שנכתבו על ידי העמיתים סווגו בקטגוריות ׳הצהרה׳ או ׳אימות׳, המוגדרות כמשוב באיכות נמוכה עד בינונית. לעומת זאת, הערות מסוג ׳הכוונה׳ - שהוגדרו כמשוב באיכות הגבוהה ביותר - היו נפוצות כמעט פי שישה בקרב הצ׳טבוטים אל מול העמיתים. ממצאים אלו עולים בקנה אחד עם מחקרים קודמים שהראו כי עמיתים נוטים לספק פחות הצעות קונקרטיות לשיפור, ובמקום זאת מתמקדים בהסבר ההיגיון שמאחורי השיפוטים שלהם (Usher & Barak, 2018; Voet et al., 2018). יתר על כן, ייתכן שהפערים הללו קשורים לתחושת חוסר ביטחון של סטודנטים או חוסר הנוחות המתלווה למתן משוב ביקורתי לעמיתיהם (Usher & Barak, 2018; Li et al., 2020).

למרות שהצ׳טבוטים סיפקו משובים מפורטים ועקביים, היו מקרים שבהם הציעו משוב שאינו תואם את ההנחיות ולעיתים אף סותר את התוכן הנלמד. מגבלה זו עולה בקנה אחד עם מחקרים שדיווחו כי מערכות בינה מלאכותית עלולות לעיתים לפרש באופן שגוי את תגובות הלומדים, מה שעשוי להוביל לחוסר דיוק במשוב (Chan & Hu, 2023; Lu et al., 2024). לעומת זאת, משובי העמיתים היו מותאמים אישית יותר, התייחסו לתוכן הקורס ולדיונים שנערכו במסגרתו, ושיקפו את נקודות המבט והניסיון הייחודיים של כל מעריך. סוג זה של משוב יכול לעודד מעורבות עמוקה יותר עם התוכן, ליצור חיבור משמעותי לחוויית הלמידה (Usher & Barak, 2018; Usher & Barak, 2024) וכן לתרום ליצירת סביבת למידה שיתופית ותומכת (Deeley & Bovill, 2017; Li et al., 2020).

לסיכום, ממצאי המחקר מדגישים את הסיכון הפוטנציאלי שבהסתמכות יתר על משובים שהופקו על ידי צ׳טבוטים, שעלולים להטעות את הסטודנטים באם לא ייבחנו וייערכו בצורה ביקורתית ומושכלת. הדבר

ממחיש ביתר שאת את הצורך בפיקוח מתמשך והשגחה מתמדת בכדי להבטיח את שילובם היעיל של כלים מבוססי בינה מלאכותית בהערכות חינוכיות (Chan & Hu, 2023; Lu et al., 2024). כמו כן, ממצאים אלו מדגישים את החשיבות שבהקניית מיומנויות אוריינות בינה מלאכותית בקרב הסטודנטים, אשר יאפשרו להם להעריך בצורה ביקורתית את איכות המשובים המופקים על ידי הצ׳טבוטים ולהשתמש במערכות אלה בצורה יעילה ומושכלת אשר תביא לתוצאות הרלוונטיות והמדויקות ביותר עבורם (Ding et al., 2023; Nikolopoulou, 2024; Walter, 2024). הממצאים מדגישים את חשיבותו של השיפוט האנושי בלכידת תובנות מעמיקות שהבינה המלאכותית עשויה לפספס, ומבליטים את הסיכונים שבהסתמכות בלעדית על מערכות אלו לצורכי הערכה. גישה מאוזנת, שמשלבת את היתרונות של צ׳אטבוטים לצד הערכות אנושיות, עשויה לספק תובנות רלוונטיות ובעלות הקשר מדויק, לצד היעילות והעקביות שמביאה עימה המכונה. באופן זה ניתן לנצל את היכולות של הבינה המלאכותית הג׳נרטיבית מבלי לוותר על המגע האישי וההבנה ההקשרית של מעריכים אנושיים. הדבר יסייע להבטיח כי הטכנולוגיה תשלים – ולא תחליף – את המרכיבים האנושיים הכה קריטיים של תהליכי ההערכה.

מקורות

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

Deeley, S. J., & Bovill, C. (2017). Staff student partnership in assessment: Enhancing assessment literacy through democratic practices. *Assessment & Evaluation in Higher Education 42*(3), 463–477. https://doi.org/10.1080/02602938.2015.1126551

Ding, L., Li, T., Jiang, S., & Gapud, A. (2023). Students' perceptions of using ChatGPT in a physics class as a virtual tutor. *International Journal of Educational Technology in Higher Education, 20*(1), 63. https://doi.org/10.1186/s41239-023-00434-1

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.

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

Nikolopoulou, K. (2024). Generative artificial intelligence in higher education: Exploring ways of harnessing pedagogical practices with the assistance of ChatGPT. *International Journal of Changes in Education, 1*(2), 103-111 .https://doi.org/10.47852/bonviewIJCE42022489

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

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

Voet, M., Gielen, M., Boelens, R., & De Wever, B. (2018). Using Feedback Requests to Actively Involve assessees in peer assessment: Effects on the assessor's feedback content and assessee's agreement with feedback. *European Journal of Psychology of Education, 33*, 145–164. https://doi.org/10.1007/s10212-017-0345-x

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. http://dx.doi.org/10.1080/02602938.2017.1405238

Usher, M. & Barak, M. (2024). Unpacking the role of AI ethics online education for science and engineering students. *International Journal of STEM Education, 11*(35). https://doi.org/10.1186/s40594-024-00493-4

Walter, Y. (2024). Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education, 21*(1), 15. https://doi.org/10.1186/s41239-024-00448-3