בין דיוק מדעי לחדשנות טכנולוגית:
הערכת תוכן שנוצר באמצעות בינה מלאכותית יוצרת
(מאמר קצר)

עידית גת
הטכניון – מכון טכנולוגי לישראל
Idit.gat@campus.technion.ac.il

מאיה אושר
HIT מכון טכנולוגי חולון
הטכניון – מכון טכנולוגי לישראל
mayau@ed.technion.ac.il

מירי ברק
הטכניון – מכון טכנולוגי לישראל
bmiriam@ed.technion.ac.il

Between scientific accuracy and technological innovation:
Evaluation of AI-generated content
(Short paper)

Idit Gat
Technion – Israel Institute of Technology
Idit.gat@campus.technion.ac.il

Maya Usher
HIT Holon Institute of Technology
Technion – Israel Institute of Technology
mayau@technion.ac.il

Miri Barak
Technion – Israel Institute of Technology
bmiriam@ed.technion.ac.il

Abstract

The growing integration of generative artificial intelligence (GenAI) technologies into science education underscores the need for systematic, critical evaluation of AI-generated content, which may contain scientific inaccuracies and reinforce misconceptions. This mixed-methods study examined how science teachers evaluate content generated by ChatGPT and how they perceive the integration of such evaluative practices within their teaching. Sixty middle-school science teachers participated in a professional development (PD) workshop at the Technion. During the workshop, teachers used ChatGPT to generate questions at different cognitive levels and corresponding answers, then evaluated them drawing on their pedagogical content knowledge for scientific accuracy, linguistic clarity, and curriculum alignment. A dual-analytic approach combined quantitative analyses of numerical ratings with qualitative analyses of written explanations and reflections. The findings revealed that lower-order thinking questions received consistently high evaluations, whereas higher-order thinking questions were perceived as containing conceptual gaps and vague formulations. Nonetheless, teachers recognized the pedagogical value of such questions for fostering reflective thinking and connecting the content to everyday contexts. Participants' reflections emphasized that structured evaluation of GenAI outputs can enhance instruction by promoting disciplinary discourse, broadening assessment practices, bridging differences in teaching experience, and redefining the teacher's role as an "evaluator." The findings highlight the importance of incorporating critical evaluation practices of AI-generated content into PD programs for science

teachers as a prerequisite for the informed and responsible integration of technological innovation in education.

Keywords: Generative artificial intelligence (GenAI), evaluation, professional development (PD), science education.

תקציר

שילובן הגובר של טכנולוגיות בינה מלאכותית יוצרת (במ"י) בהוראת המדעים מחייב פיתוח מיומנויות הערכה ביקורתיות ושיטתיות, שכן תוכן הנוצר באמצעותן עלול לכלול טעויות מדעיות ולהנציח תפיסות שגויות. מחקר אמפירי זה בחן כיצד מורים למדעים מעריכים תוכן שנוצר על ידי במ"י וכיצד הם תופסים את שילובן של פרקטיקות הערכה אלו במסגרת הוראתם. משתתפי המחקר כללו שישים מורים בחטיבות ביניים אשר לקחו חלק בסדנה לפיתוח מקצועי בטכניון. במסגרת הסדנה, המורים יצרו באמצעות ChatGPT שאלות מסדר חשיבה שונה ותשובות תואמות, והעריכו אותן בהתבסס על הידע הפדגוגי-מקצועי שברשותם, לפי קריטריונים של דיוק מדעי, בהירות לשונית והתאמה לתוכנית הלימודים. הנתונים נותחו בגישה דואלית-אנליטית, שכללה ניתוח כמותי של הדירוגים המספריים לצד ניתוח איכותני של ההסברים והרפלקציות שהוגשו. ממצאי המחקר חשפו כי שאלות מסדר חשיבה נמוך זכו להערכות גבוהות ועקביות, בעוד ששאלות מסדר חשיבה גבוה נתפסו כמציגות פערים מושגיים וניסוחים מעורפלים. עם זאת, המורים זיהו את תרומתן הפדגוגית של שאלות אלו לטיפוח חשיבה רפלקטיבית ולקישור החומר הנלמד להקשרים יומיומיים. רפלקציות המשתתפים הדגישו כי הערכה ביקורתית של תוצרי במ"י תורמת ישירות לשיפור איכות ההוראה באמצעות חיזוק השיח הדיסציפלינרי, שכלול גישות הערכה חדשניות, גישור על פערי נסיון בהוראה, והגדרה מחודשת של תפקיד המורה כ"מעריך". הממצאים מדגישים את חשיבותן של שילוב פרקטיקות להערכה ביקורתית של תוצרי במ"י בתוכניות פיתוח מקצועי למורי מדעים, כתנאי לשילוב מושכל ואחראי של חדשנות טכנולוגית בחינוך.

מילות מפתח: בינה מלאכותית יוצרת (במ"י), הערכה, פיתוח מקצועי, הוראת מדעים.

מבוא

שילובה של בינה מלאכותית יוצרת (במ"י) במערכת החינוך טומן פוטנציאל חינוכי משמעותי, אך במקביל מציב אתגרים פדגוגיים ואפיסטמולוגיים (Adiguzel et al., 2023; Chang et al., 2024). אתגרים אלו מתחדדים במיוחד בהקשר של החינוך המדעי, שבו דיוק מושגי וחשיבה דיסציפלינרית הם תנאים הכרחיים להבנה ולפיתוח אוריינות מדעית (Usher & Barak, 2025; Blonder et al., 2024). מודלים מבוססי שפה (LLMs) כגון ChatGPT מסוגלים להפיק טקסטים רהוטים ומשכנעים, אולם לעיתים עושים זאת על חשבון הדיוק העובדתי, תופעה המכונה "הזיות" (Birhane et al., 2023).

הספרות מצביעה על מספר ממדים מרכזיים להערכת תוצרים שנוצרו ע"י במ"י: דיוק מדעי, בהירות לשונית והתאמה לתוכנית הלימודים. דיוק מדעי מהווה מימד מרכזי, במיוחד בהקשרים של חינוך מדעי. במחקרים אחרונים נמצא כי במבחנים ארציים מערכות אלו הפגינו לעיתים קרובות ביצועים חלקיים או שגויים, ואף עשויות לחזק תפיסות שגויות בקרב לומדים (Feldman-Maggor et al., 2025; Yamtinah et al., 2025). מימד נוסף, בהירות לשונית, מוגדר בספרות כמאפייניו של טקסט תקין תחבירית, מדויק ונגיש לקהל היעד. מחקרים מצביעים על יכולתן של מערכות במ"י להפיק תוכן רהוט ובעל ערך פדגוגי (Mahowald et al., 2024; Lee et al., 2024). בנוסף, בוזיאן ובוזיאן (2024) מצאו כי ChatGPT יעיל במיוחד בתיקון שגיאות טכניות ובהפקת ניסוחים תקינים, אך מוגבל ביכולתו להעריך עומק מושגי. באשר להתאמה לתוכנית הלימודים, ממד זה טרם נבחן לעומק. עדויות עדכניות מצביעות על כך שרמת ההתאמה תלויה במידה רבה בתיווך פעיל של מורים, אשר מרחיבים ומתאימים את הפלטים של הבמ"י על בסיס הידע הדיסציפלינרי והניסיון הפדגוגי (Karataş et al., 2025). למרות ששלושת הממדים הללו, מוכרים כחיוניים להערכת תוכן שנוצר על ידי במ"י, מרבית המחקרים בחנו כל ממד בנפרד, ולא התעמקו באופן שבו מורים משלבים אותם בתהליכי הערכה בפועל.

בנוסף, הספרות מתארת כי מרבית תוכניות הפיתוח המקצועי הקיימות למורים מתמקדות בהיבטי אוריינות טכנולוגית או ביישומים פדגוגיים (Chiu, 2025; Ng et al., 2023), בעוד שטיפוח מיומנויות הערכה ביקורתיות נמצא בשלבים ראשוניים.

פערים אלו מדגישים את הצורך בביצוע מחקרים אמפיריים שיבחנו כיצד מורים מעריכים בפועל תוצרים מדעיים שנוצרו על ידי במ"י, וכיצד הם תופסים את שילובן של פרקטיקות הערכה אלו בהוראה.

שאלות המחקר

המחקר הונחה על ידי שתי שאלות מחקר:

1. כיצד מורים למדעים מעריכים שאלות ותשובות שנוצרו באמצעות במ"י, מבחינת דיוק מדעי, בהירות לשונית והתאמה לתוכנית הלימודים?
2. כיצד תופסים המורים את שילובן של פרקטיקות הערכה מבוססות במ"י בהוראת המדעים?

משתתפים

במחקר השתתפו 60 מורים למדעים מחטיבות-ביניים. 55% הוגדרו כמורים מתחילים (בעלי פחות מחמש שנות ניסיון) ו־ 45% כמורים ותיקים (בעלי חמש שנות ניסיון ומעלה). המשתתפים ייצגו דיסציפלינות שונות: פיזיקה (28%), כימיה (26%), ביולוגיה (22%), מתמטיקה (12%) ומדעי המחשב (12%). 75% מהם העידו כי הם משלבים במ"י בהוראה.

מהלך המחקר

המחקר נערך בטכניון בשנת תשפ"ה, במסגרת סמינר פיתוח מקצועי שעסק בשילוב במ"י בהוראת המדעים וכלל הרצאה, סדנה מעשית ודיון. המשתתפים חולקו לקבוצות הטרוגניות (מורים ותיקים וחדשים), לפי תחומי דעת. בסדנה המעשית, שהיוותה את ליבת הסמינר, הזינו הקבוצות מאמר מדעי ל-ChatGPT. באמצעות הנחיה (Prompt) אחידה, הם ביקשו מהמערכת לנסח שתי שאלות פתוחות: אחת ברמת חשיבה נמוכה (הבנה) ואחת ברמת חשיבה גבוהה (בגון ניתוח או יצירה). לאחר מכן, העריכו המורים את איכות השאלות באמצעות הידע תוכן המקצועי והפדגוגי שלהם, בחרו אחת מהן, הפיקו עבורה תשובה באמצעות הצ'אט וביצעו הערכה זהה גם לתוכן התשובה. לבסוף, ענו הקבוצות על שאלת רפלקציה שבחנה את חוויית הלמידה ואת תפיסותיהם בנוגע להערכת תוצרי במ"י בהוראה.

שיטה וכלי מחקר

המחקר יישם גישה אנליטית משולבת (כמותית ואיכותנית) לניתוח מערך נתונים יחיד (Bazeley, 2012). נעשה שימוש בשני כלים לאיסוף הנתונים אשר פותחו במיוחד לצורך המחקר:

1. **טופס להערכה שיטתית של תוכן שנוצר באמצעות במ"י** – הטופס פותח במטרה לענות על שאלת המחקר הראשונה, וכלל שלושה קריטריונים מרכזיים להערכת תוצרים שנוצרו על ידי במ"י: דיוק מדעי, בהירות לשונית והתאמה לתכנית הלימודים (Feldman Maggor et al., 2025; Mason, 2000; Shulman, 1986). להבטחת תוקף התוכן, נוסח הטופס נבחן על ידי שני מומחים בתחום החינוך המדעי, אשר אישרו כי הקריטריונים שנבחרו רלוונטיים, מקיפים ומנוסחים באופן ברור. כל קריטריון הוערך באמצעות סולם ליקרט בן חמש רמות ובנימוק מילולי (ראה טבלה 1).

טבלה 1. טופס להערכת תוכן שנוצר באמצעות במ"י

קריטריונים להערכה	**הדביקו כאן את נוסח השאלה/תשובה שיוצרה על ידי ChatGPT** **1- רמה נמוכה מאוד 2- רמה נמוכה 3- רמה בינונית** **4- רמה טובה 5- רמה טובה מאוד**
מידת הדיוק המדעי האם השאלה/תשובה נשענת על עקרונות מדעיים? האם המושגים שבה נכונים ומבוססים על ידע	הערכה מספרית: 1-2-3-4-5 הסבר כתוב להערכה: _____ _____

קריטריונים להערכה	**הדביקו כאן את נוסח השאלה/תשובה שיוצרה על ידי ChatGPT** **1- רמה נמוכה מאוד 2- רמה נמוכה 3- רמה בינונית** **4- רמה טובה 5- רמה טובה מאוד**
מדעי מוכר, ללא שגיאות עובדתיות?	
מידת הבהירות הלשונית עד כמה השאלה/תשובה מנוסחת באופן ברור, מובנת, ונכונה תחבירית?	הערכה מספרית: 1-2-3-4-5 הסבר כתוב להערכה: _____ _____
מידת התאמה לתכנית הלימודים האם השאלה/תשובה מותאמת מבחינת רמת תוכן, שפה ואוצר מילים לתלמידים בחטיבת-ביניים?	הערכה מספרית: 1-2-3-4-5 הסבר כתוב להערכה: _____ _____

2. **טופס רפלקציה** – נבנה בכדי לענות על שאלת המחקר השנייה ובחן כיצד המורים תופסים שילוב תהליכי הערכה של תוכן שנוצר באמצעות במ"י בהוראה.

בסיום הסדנה נאספו 15 תוצרים קבוצתיים (שאלות, תשובות ורפלקציות), אשר שימשו כבסיס לניתוח המחקרי.

ניתוח הנתונים

ניתוח הנתונים שילב גישות כמותיות ואיכותניות. בפן הכמותי, עובדו דירוגי המשתתפים באמצעות מבחן וילקוקסון למדגמים תלויים וניתוח בוטסטראפ בייסיאני (Bayesian bootstrap) להערכת מובהקות ההבדלים ועוצמתם. בפן האיכותני, נימוקי ההערכה נותחו בשיטה דדוקטיבית, והרפלקציות בשיטה אינדוקטיבית.(Hsieh & Shannon, 2005) מהימנות בין שופטים, שנבחנה על ידי שני מומחים, הניבה מדד קאפה של כהן של 0.91.

ממצאים

1. הערכות המורים את השאלות והתשובות שהופקו ע"י במ"י

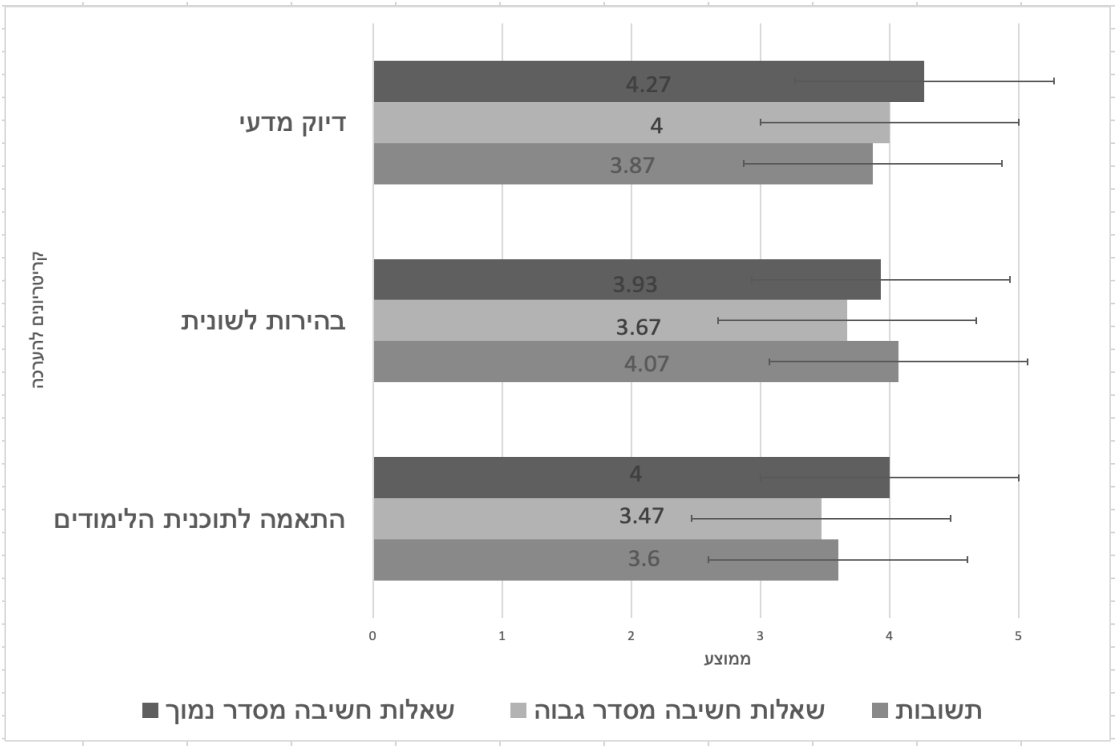
ניתוח הערכות המורים לשאלות החשיבה מסדר נמוך – *דיוק מדעי* קיבל את הציון הגבוה ביותר ביחס לשני הקריטריונים האחרים (M = 4.27, SD = 0.46). *בהירות לשונית* דורג מעט נמוך יותר (M = 3.93, SD = 0.70), באשר המורים ציינו כי ישנם מונחים שעשויים לדרוש הבהרה: "*השאלה מנוסחת באופן תקין מבחינה דקדוקית ומבנית, אך ביטויים כגון 'מודל חלקיקי' יידרשו הסבר*" (פיזיקה-קב.7). *הלימה לתוכנית הלימודים* דורגה באופן חיובי (M = 4, SD = 0.65).

ניתוח הערכות המורים לשאלות חשיבה מסדר גבוה – דיוק מדעי דורג באופן חיובי (M = 4, SD = 0.65), לדוגמה, מורי פיזיקה-קב.8 סברו: "*השאלה עוסקת היטב במושגים מרכזיים כגון אינרציה, התנגדות אוויר ואנרגיה...השאלה מדויקת ובעלת היגיון פיזיקלי*". בהירות לשונית דורג נמוך יותר (M = 3.67, SD = 0.82), והלימה לתוכנית הלימודים דורג הכי נמוך (M = 3.47, SD =0.92), באשר סטיית התקן הגבוהה יחסית מצביעה על פערי תפיסות בקרב המורים. חלקם מצאו את השאלות "*מתאימות לרמת הגיל*", בעוד אחרים סברו שהשאלה חורגת מרמת המוכנות הקוגניטיבית של תלמידי חטיבת-ביניים.

באשר לתשובות, המורים העניקו להן דירוגים חיוביים, אם כי מתונים בהשוואה לשאלות. דיוק מדעי דורג באופן חיובי (M = 3.87, SD = 0.52), בהירות לשונית דורג גבוה (M = 4.07, SD = 0.70), התשובות נתפסו כמנוסחות "*באופן ברור וקריא*", גם כאשר היו מורכבות יותר. הלימה לתוכנית הלימודים דורג בינוני (M = 3.6, SD = 0.83), וסטיית התקן הגבוהה הצביעה על פערי תפיסות. לדוגמה מורי ביולוגיה-קב.13 ראו בתשובות: "*העמקה מועילה המאפשרת הרחבת ידע*", בעוד מורי מתמטיקה-קב.5 תארו: "המגבלה העיקרית שזיהינו היא חוסר ההתאמה לרמת הכיתה [..] *ועלולה ליצור עומס קוגניטיבי*".

השוואת הדירוגים באמצעות מבחני וילקוקסון וניתוח בייסיאניי**,** הצביעה על יתרון עקבי לשאלות ברמת חשיבה נמוכה על שאלות ברמת חשיבה גבוהה במובהקות ובעוצמה.

איור 1 מציג את הממוצעים וסטיות התקן של הערכות המורים את השאלות והתשובות שנוצרו באמצעות במ"י בחלוקה לשלושת הקריטריונים.



איור 1. ממוצעים וסטיות תקן של הערכות המורים את השאלות והתשובות שנוצרו באמצעות במ"י

2. תפיסות מורים באשר לשילוב פרקטיקות הערכה מבוססות במ"י בחינוך המדעי

הניתוח האיכותני של הרפלקציות הניב 46 מקטעי טקסט, אשר אורגנו לארבע קטגוריות מרכזיות:

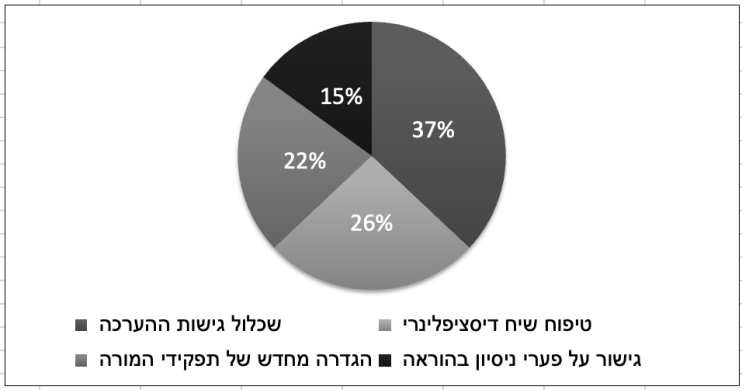
שכלול גישות ההערכה (37%): המורים תופסים את תוצרי הבמ"י כ"נקודת התחלה, לא תוצר סופי", הדורשת עיבוד, אימות והתאמה למטרות ההוראה. התהליך תואר כ"תרגיל בהערכה ביקורתית" המעודד דיוק מושגי ולשוני.

טיפוח שיח דיסציפלינרי (26%): ההערכה הקבוצתית תרמה להעמקת השיח המקצועי ולפיתוח שפה משותפת. הדיון בצוות אפשר לזהות חוזקות וחולשות בתוכן, להבין מה מבלבל בשאלה וכיצד היא תואמת לתוכנית הלימודים.

הגדרה מחדש של תפקידי המורה (22%): ניכר מעבר מתפקיד של משתמש פסיבי למתווך בעל אחריות מקצועית: "אנחנו לא רק משתמשים בכלי, אנחנו אחראים למה שהוא מייצר". המורים הדגישו את הצורך בפיקוח על הפלט ובחינוך תלמידים להבחנה בין מידע מהימן לשגוי.

גישור על פערי ניסיון (15%): העבודה המשותפת אפשרה למידה הדדית; מורים חדשים רכשו ביטחון ומיומנויות הערכה ממורים מנוסים, בעוד שהוותיקים נהנו מפרספקטיבות רעננות ומ"שאלות שהוותיקים לא היו חושבים עליהן".

איור 2 מתאר את התפלגות תפיסות מורים לארבע קטגוריות מרכזיות.



איור 2. התפלגות תפיסות המורים באשר לשילוב פרקטיקות הערכה מבוססות במ"י בהוראה

דיון

ממצאי המחקר הראו כי שאלות ברמת חשיבה נמוכה דורגו גבוה יותר בדיוק מדעי, בהירות והתאמה, וזאת בהלימה לספרות המצביעה על ירידה באמינות תוצרי במ"י ככל שהמורכבות עולה (Thanh et al., 2023; Feldman-Maggor et al., 2025; Yamtinah et al., 2025). המורים תפסו את הפלטים בנקודת מוצא הדורשת עיבוד ביקורתי, ממצא המדגיש את חשיבות התיווך האנושי להבטחת תקפות הידע (Arantes, 2024; Erduran, 2025). אף שהבהירות הלשונית נמצאה בגבוהה יחסית, הוכח כי היא אינה ערובה לנגישות קוגניטיבית (Mahowald et al., 2024; Zhai et al., 2025). כמו כן, שאלות מורכבות נטו לחרוג מתוכנית הלימודים, דבר המחזק את הספרות אודות תפקיד המורה כמתווך (Mason, 2000).

ממצא מרכזי חושף כי העבודה השיתופית בסדנה תרמה לשכלול שיטות ההערכה ולקידום הערכה מושכלת המבוססת על אימות, דיוק ודיון רפלקטיבי בעת שילוב תוצרי במ"י. בנוסף, הסדנה תרמה להתפתחותם המקצועית של המשתתפים באמצעות העמקת השיח הפדגוגי ויצירת מרחב ללמידה הדדית, שבו פעלו המורים כמעצבים ביקורתיים של ידע, ולא כמשתמשים פסיביים בטכנולוגיה. לבסוף, נמצא כי ההתנסות תרמה באופן שונה למורים מתחילים ומנוסים: המתחילים חיזקו את ביטחונם ובישורי ההערכה שלהם, בעוד שהמנוסים נהנו מפרספקטיבות חדשות ופתיחות לרעיונות חדשניים. ממצא זה מתיישב עם מודלים של למידה מקצועית המדגישים את הערך של אינטראקציה בין בעלי ניסיון שונה (Chen et al., 2025; Vangrieken et al., 2016; Rodgers & Skelton, 2014).

לסיכום, במ"י מהווה כלי עזר להרחבת משאבי ההוראה אך אינה תחליף לשיקול דעת פדגוגי. המחקר מציע פרקטיקה שיטתית להערכת תוצרי במ"י וקובע כי שילובה בהכשרת מורים עשויה לסייע ליישום טכנולוגי אחראי ואפקטיבי במערכת החינוך.

מקורות

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology, 15*(3). https://doi.org/10.30935/cedtech/13152

Arantes, J. (2024). Understanding intersections between GenAI and pre-service teacher education: What do we need to understand about the changing face of truth in science education?. *Journal of Science Education and Technology*, 1-12. https://doi.org/10.1007/s10956-024-10189-7

Bazeley, P. (2012). Integrative analysis strategies for mixed data sources. *American Behavioral Scientist, 56*(6), 814-828.

Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, 5(5), 277-280. https://doi.org/10.1038/s42254-023-00581-4

Blonder, R., Feldman-Maggor, Y., & Rap, S. (2024). Are they ready to teach? Generative AI as a means to uncover pre-service science teachers' PCK and enhance their preparation program. *Journal of Science Education and Technology*, 1-10. https://doi.org/10.1007/s10956-024-10180-2

Chen, R., Lee, V. R., & G Lee, M. (2025). A cross-sectional look at teacher reactions, worries, and professional development needs related to generative AI in an urban school district. *Education and Information Technologies*, 1-38. https://doi.org/10.1007/s10639-025-13350-w

Chiu, T. K. (2025). Developing intelligent-TPACK (I-TPACK) framework from unpacking AI literacy and competency: implementation strategies and future research direction. *Interactive Learning Environments*, 33(7), 4189–4192. https://doi.org/10.1080/10494820.2025.2545053

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on the evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1-45.

Feldman-Maggor, Y., Blonder, R., & Alexandron, G. (2025). Perspectives of generative AI in chemistry education within the TPACK framework. *Journal of Science Education and Technology*, 34(1), 1-12. https://doi.org/10.1007/s10956-024-10147-3

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. Qualitative Health Research, 15(9), 1277-1288.

Kohen-Vacs, D., Usher, M. & Jansen, M. Integrating Generative AI into Programming Education: Student Perceptions and the Challenge of Correcting AI Errors. *Int J Artif Intell Educ* **35**, 3166-3184 (2025). https://doi.org/10.1007/s40593-025-00496-4

Karataş, F., Eriçok, B., & Tanrikulu, L. (2025). Reshaping curriculum adaptation in the age of artificial intelligence: Mapping teachers' AI-driven curriculum adaptation patterns. *British Educational Research Journal*, 51(1), 154-180. https://doi.org/10.1002/berj.4068

Lee, S., & Song, K. S. (2024). Teachers' and students' perceptions of AI-generated concept explanations: Implications for integrating generative AI in computer science education. *Computers and Education: Artificial Intelligence*, 7, 100283.

Lan, Y. J., & Chen, N. S. (2024). Teachers' agency in the era of LLM and generative AI. *Educational Technology & Society,* 27(1), I-XVIII. https://www.jstor.org/stable/48754837

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6), 517-540.

Mason, M. (2000). Teachers as critical mediators of knowledge. *Journal of Philosophy of Education*, 34(2), 343-352.

Ng, D. T. K., Leung, J. K. L., Su, J., Ng, R. C. W., & Chu, S. K. W. (2023). Teachers' AI digital competencies and twenty-first century skills in the post-pandemic world. *Educational Technology Research and Development*, 71(1), 137-161.

Rodgers, C., & Skelton, J. (2014). Professional Development and Mentoring in Support of Teacher Retention. *Journal on School Educational Technology*, 9(3), 1-11.

Shulman, L. S. (1986). Those who understand: A conception of teacher knowledge. *American Educator*, 10(1).

Thanh, B. N., Vo, D. T. H., Nhat, M. N., Pham, T. T. T., Trung, H. T., & Xuan, S. H. (2023). Race with the machines: Assessing the capability of generative AI in solving authentic assessments. *Australasian Journal of Educational Technology*, 39(5), 59-81. https://doi.org/10.14742/ajet.8902

Usher, M., Barak, M. & Erduran, S. What role should higher education institutions play in fostering AI ethics? Insights from science and engineering graduate students. *IJ STEM Ed* **12**, 51 (2025). https://doi.org/10.1186/s40594-025-00567-x

Vangrieken, K., Dochy, F., & Raes, E. (2016). Team learning in teacher teams: Team entitativity as a bridge between teams-in-theory and teams-in-practice. *European Journal of Psychology of Education*, 31(3), 275-298.

Zhai, X., Nyaaba, M., & Ma, W. (2025). Can generative AI and ChatGPT outperform humans on cognitive-demanding problem-solving tasks in science?. *Science & Education*, 34(2), 649-670. https://doi.org/10.1007/s11191-024-00496-1