השפעת הנראות של אוואטאר מבוסס Chat-GPT על האפקטיביות של אימון לראיונות עבודה

דניאל ז׳ורבל
המכללה האקדמית להנדסה בראודה בכרמיאל
daniel.zhuravel@e.braude.ac.il

נירית גביש
המכללה האקדמית להנדסה בראודה בכרמיאל
nirit@braude.ac.il

איציק בן-שלוש
Holistic EHS / XR
itzik@holisticehs.com

The Effect of a ChatGPT-Based Avatar's Appearance on the Effectiveness of Job Interview Training

Daniel Zhuravel
Braude College of Engineering, Karmiel
daniel.zhuravel@e.braude.ac.il

Nirit Gavish
Braude College of Engineering, Karmiel
nirit@braude.ac.il

Itzik Ben-Shlush
Holistic EHS / XR
itzik@holisticehs.com

Abstract

We examined training for a job interview through a role-play simulation using a ChatGPT-based chatbot. We questioned whether a humanized large language model (LLM)-based chatbot avatar that mimics human-to-human interaction conventions through its appearance – i.e., a formal-looking chatbot who plays the role of the interviewer – will improve the effectiveness and the user's subjective evaluation of the training. Three avatars, each visually different, were used: formal, sporty, and casual avatar. The 85 trainees were randomly assigned to one of the three groups, and each received job interview training by the assigned avatar as the interviewer. Trainees received an evaluation of their performance during the interview and then repeated the job interview simulation again. The results demonstrate that the Casual Avatar group spent less time training than the Sporty and Formal Avatar groups. All the groups improved in the second session compared to the first, but the improvement did not differ from group to group. Trainees' answers to the subjective evaluation questions did not indicate a significant difference between the groups. We conclude that in LLM-based role play with chatbots for training social skills, the avatar's appearance should probably mimic real-life conventions.

Keywords: LLM; ChatGPT; chatbot; training; job interview; role-play.

תקציר

המחקר הנוכחי בחן את האפקטיביות של אימון לראיון עבודה באמצעות סימולציית משחק תפקידים במחשב, בה השתמשנו בצ׳אטבוט מבוסס ChatGPT הנראה בצורות שונות. רצינו להבין האם השימוש בצ׳אטבוט המבוסס מודלי שפה גדולים (Large Language Models – LLMs) שמראהו תואם למראה קונבנציונלי של מראיין אנושי ישפר את אפקטיביות האימון ואת שביעות הרצון של המשתמש בנוגע לאימון שעבר יחסית למראה שאינו אופייני למראיין. לצורך כך, בחרנו שלושה אוואטארים, שלכל אחד מראה אחר: פורמלי, ספורטיבי ויום-יומי. 85 מתאמנים שובצו לאחד משלושת האוואטארים באופן אקראי, ואותו האוואטאר שימש עבורם כמראיין באימון לראיון עבודה. המשתתפים ביצעו סימולציה של ראיון עבודה עם האוואטאר, קיבלו הערכת ביצועים בסוף הראיון, ולאחריו ביצעו את הסימולציה מחדש. תוצאות המחקר הראו כי הקבוצה שביצעה את האימון עם האוואטאר במראה היום-יומי עשתה זאת בפחות זמן

לעומת הקבוצות שהתאמנו עם האוואטאר הפורמלי והספורטיבי. כל הקבוצות השתפרו בין הראיון הראשון לשני, אך לא היה שוני ברמת השיפור בין הקבוצות. בסוף התהליך קיבלו המשתתפים שאלון שבו נשאלו שאלות שביעות הרצון, ואלו לא הראו על שוני מובהק בין הקבוצות. לסיכום, מהתוצאות ניתן להסיק כי בעת שימוש בצ׳אטבוט המבוסס LLM עבור אימונים לשיפור יכולות חברתיות באמצעות משחק תפקידים האוואטאר צריך לדמות במראהו ייצוג אנושי מתאים לחיים האמיתיים.

מילות מפתח: ChatGPT, LLM, צ׳אטבוט, ראיון עבודה, משחק תפקידים.

מבוא

סקירת ספרות

התפתחות הבינה המלאכותית (AI) הביאה לעלייה בשימוש בצ׳אטבוטים. צ׳אטבוטים מהווים סוכנים ממוחשבים שיכולים לקיים אינטרקציה עם משתמשים תוך שימוש בשפה טבעית באמצעות קול (דיבור) או טקסט (Huang et al., 2007). כאשר פותחו אלגוריתמי עיבוד השפה הטבעית המבוססים למידה עמוקה, תפקיד הצ׳אטבוט השתנה באופן משמעותי, וכיום נוגע כמעט בכל תחום של חיינו המקצועיים והיום-יומיים. אלגוריתמים אלו נקראים מודלי שפה גדולים (Large Language Models – LLMs) והם הבסיס לצ׳אטבוטים חזקים מאוד כמו ChatGPT (OpenAI, 2022), Claude (Claude.AI, 2024), Microsoft Copilot Microsoft, (2024), ואחרים. צ׳אטבוטים המבוססים LLM מסוגלים לנהל שיחות טבעיות עם בני אדם, להבין ולהגיב באופן דמוי אנושי, הם רגישים להקשר, הם יכולים לייצר מידע חדש, הם מדוייקים והם לא מוגבלים לתחום ספציפי (Dam et al., 2024; Wan, 2024).

לכן, זה לא מפתיע שיש שימוש בצ׳אטבוטים המבוססים LLM במגוון רחב של תחומים ומשימות, רבים מהם הינם תחומים חדשים בעבור צ׳אטבוטים, שטרם נחקרו באופן מעמיק. תחום מעניין בו נעשה שימוש בצ׳אטבוטים מבוססי LLM הוא סימולציות משחקי תפקידים לאימון כישורים חברתיים. כישורים חברתיים מוגדרים כיכולת לשלוח ולקבל מידע רגשי, יכולות הבעה עצמית בצורה מילולית, הבנת מסרים ומידע מילולי, וויסות תקשורת רגשית והתנהגות מילולית ופיתוח מיומנויות הצגה עצמית (Riggio, 1986). שיטה מומלצת לשימוש בצ׳אטבוט מבוסס LLM באימון כישורים חברתיים היא לתת לצ׳אטבוט תפקיד כפול – הן כשותף המשחק בתפקיד והן כמנטור המנתח את האינטראקציה ומספק משוב, בדרך כלל על בסיס תיאוריות ומתודולוגיות ידועות בתחום הספציפי בו מתקיים האימון (Yang et al., 2024).

השימוש בצ׳אטבוט לצורך סימולציית משחק תפקידים לאימון כישורים חברתיים מדגיש את החשיבות בצורך לגרום למשתמש לשתף פעולה עם הצ׳אטבוט ולבטוח בו. דרך אחת להשיג זאת היא באמצעות צ׳אטבוט המבוסס על דמות אוואטאר. על פי גישת ״מחשבים הם שחקנים חברתיים״ (CASA, Reeves & Nass, 1996), משתמשים מתייחסים לטכנולוגיות כאל אנשים אמיתיים באופן לא מודע. בהתאם לכך, הממשקים צריכים צריכים להזכיר דמות אנושית ככל הניתן, כיוון שאז הם ידרשו פחות התאמה מצד המשתמשים לצורך האינטראקציה החברתית המתבצעת איתם. גישה זו נתמכת על ידי ממצאים המראים שדמויות המזכירות אדם משפרות שיתוף פעולה (Parise et al., 1999), את האחריות להשלמת משימות בצוות משותף של אדם ומכונה (Hinds, 2004), ואת הקשר עם הדמויות (Lee et al., 2005). בהתאם לכך, ממשקים חדשים נוספו לאחרונה לממשקי טקסט מוכרים של צ׳אטבוטים מבוססי LLM. בממשקים אלו, הצ׳אטבוט מיוצג כאוואטאר בעל מראה אנושי (Sonlu et al., 2024; Wang et al., 2024; Yamazaki et al., 2023), כשהאינטראקציה יכולה להתרחש בסביבה וירטואלית או במציאות רבודה (Llanes-Jurado et al., 2024; Wan et al., 2024).

בהתבסס על האמור לעיל, עולה השאלה, האם על מנת לשפר את האמון ושיתוף הפעולה של המשתמש, יש לגרום לצ׳אטבוט המבוסס על אוואטאר דמוי-אדם לחקות במראהו את מוסכמות האינטראקציה בין בני אדם. למרבה הצער, המחקר על השפעת מראה האווטארים של צ׳אטבוטים על המשתמשים הוא דל. Wang et al. (2023) מראים שכדי לשפר את כוונת המשתמש להמשיך באינטראקציה עם הצ׳אטבוט, יש להתאים את לבושו של מדריך טיולים מבוסס AI לסגנון השיחה שלו – לבוש פורמלי עבור סגנון המשדר סמכותיות ולבוש לא פורמלי עבור סגנון המשדר חום. בניסוי נוסף, כאשר סגנון הלבוש של הצ׳אטבוטים תאם את הקונבנציות של המוצר שהם מייצגים (לדוגמה, צ׳אטבוטים של אלקטרוניקה וטיפוח גוף התלבשו באופן פורמלי, צ׳אטבוט של ביגוד ספורט התאפיין במראה אנרגטי, וצ׳אטבוט מומחה אופנה הופיע בלבוש ותסרוקת אופנתיים), חלה עלייה באמון בפלטפורמה ובכוונות רכישה (Liew et al., 2021). השאלה לגבי השפעת מראה האוואטאר חשובה במיוחד כאשר הצ׳אטבוט מיועד לסימולציות משחקי תפקידים לאימון כישורים חברתיים, כי הצלחת הסימולציה תלויה בטיב שיתוף הפעולה של המתאמן עם הצ׳אטבוט.

שאלות והשערות המחקר

במחקר הנוכחי התמקדנו באימון לראיון עבודה באמצעות סימולציית משחק תפקידים עם צ׳אטבוט מבוסס ChatGPT. שאלת המחקר הייתה האם פורמליות המראה של הצ׳אטבוט המראיין משפיעה על יעילות האימון לראיונות עבודה. בחנו את שאלת המחקר באמצעות המדדים הללו: נכונות המתאמנים להשתמש במערכת, ביצועיהם במהלך השימוש בה, והשיפור שלהם לאורך האימון. שלושה אוואטארים שונים עוצבו לשימוש כמראיינים: אוואטאר פורמלי – לבוש בחליפה; אוואטאר ספורטיבי – לבוש בבגדי ספורט; ואוואטאר במראה יום-יומי – לבוש בחולצה צבעונית ובג׳ינס. כדי למנוע שונות הנובעת ממגדר האוואטאר, כל שלושת האוואטארים עוצבו כגברים. שיערנו כי נכונות המתאמנים להשתמש במערכת, ביצועיהם ושיפורם יהיו הגבוהים ביותר עבור האוואטאר הפורמלי.

שיטת המחקר

אוכלוסייה

בניסוי השתתפו סטודנטים ממגוון רחב של חוגי ההנדסה במכללה האקדמית להנדסה בראודה בכרמיאל בעלי רמת אנגלית טובה, מכיוון שהאינטראקציה עם המערכת התנהלה באנגלית (סיימו קורס של אנגלית מתקדמים ב׳ או בעלי פטור מקורסי השלמה באנגלית). הסטודנטים נרשמו למחקר מתוך בחירה וקיבלו על השתתפותם תגמול של 100 ש״ח בתווי BuyMe. תשעת הסטודנטים בעלי הביצועים הטובים ביותר בראיונות העבודה קיבלו בונוס של 50 שקלים נוספים בתו נוסף.

בניסויים השתתפו בסה״כ 102 משתתפים, כאשר התחשבנו בתוצאות של 85 משתתפים עקב תקלות טכניות. מתוכם, השתתפו בסה״כ 60 גברים, 41 נשים, 1 שלא היה/הייתה מעוניין/ת לשתף במינם ו–1 שסימן/ה ״אחר״ בסעיף המגדר. מתוכם, התוצאות התקינות היו של 50 גברים, 33 נשים ושל השניים הנוספים. המשתתפים היו בגילאים 22-41. הגיל הממוצע היה 27.2.

29 משתתפים (18 גברים ו-11 נשים) הוקצו באופן רנדומלי לקבוצת האוואטאר הפורמלי, 28 (18 גברים, 8 נשים, 1 שלא היה/הייתה מעוניין/ת לשתף במינם ו–1 שסימן/ה ״אחר״ באפשרויות המגדר) הוקצו לקבוצת האוואטאר הספורטיבי, ו-29 (14 גברים ו-18 נשים) לקבוצת האוואטאר בלבוש היום-יומי.

כלי המחקר

לצורך המחקר השתמשנו בתכנת EON-XR (https://core.eon-xr.com/Library/Index_V2) שפותחה במטרה ליצור סביבת התממשקות עם AI המבוסס על ChatGPT באמצעות אוואטארים, בינה מלאכותית ומציאות מדומה. במערכת ניתן לעצב אוואטארים ייחודיים עם מגוון רחב של אפשרויות לעיצוב השיער, הפנים, הגוף, הלבוש, ואפילו הקול והמבטא. בנוסף, ניתן גם לעצב את האישיות של האוואטאר מדרכי התנהגות ועד לידע מקצועי. לאוואטאר ישנה היכולת להציג רשימת נושאים רלוונטיים, להרצות עליהם ולהפנות לסרטונים מתאימים שאותם הוא מכיר שיכולים להרחיב את הידע הנדרש בתחום הנלמד. בנוסף, ניתן לבצע משחק תפקידים עם האוואטאר. התקשורת עם הבינה המלאכותית מתבצעת באמצעות הקלטת פלט רצוי או הקלדתו. הקלט המתקבל מהקלטה באמצעות מיקרופון מומר לפרומפט טקסט, הבינה המלאכותית קולטת את הפרומפט ויודעת להגיב אליו וליצור תשובה רלוונטית אליו בקול שמזכיר מאוד קול אנושי.

מערך המחקר

במחקר שלנו השתמשנו במערכת EON-XR על מנת לבנות סימולציה של ראיון עבודה לתפקיד במוקד שירות לקוחות טלפוני, כאשר בסימולציה שני תפקידים: התפקיד הראשון הוא המראיין (האוואטאר) והתפקיד השני הוא המרואיין (המשתתף בניסוי). לשם הניסוי בנינו ארבעה אוואטארים עיקריים – אחד שמנחה את הסימולציה ושלושה נוספים שמראיינים את המשתמשים בזמן האימון, כשלכל אחד משלושת האחרונים שובצו המשתתפים באקראי. האוואטאר המנחה קיבל את השם ארתור, והוא מופיע באיור 1.

איור 1. האוואטאר ארתור.

ארתור עוצב כלובש חולצה מכופתרת ועניבה, בעל תספורת פורמלית, זקן מעוצב, משקפיים ומכנסיים אלגנטיים. המבטא שלו נבחר כ"בריטי סטנדרטי".

שלושת האוואטארים המראיינים קיבלו את השם מרקוס, ותיאור אישיותם היה: "מרקוס הוא מראיין פורמלי, מנומס ומנוסה. הוא יודע בדיוק מה הוא מחפש בעובדיו החדשים והוא לא מפחד להיות ישיר".

מרקוס הפורמלי, המופיע באיור 2, עוצב כלובש חליפה עם עניבה ובעל זקן ארוך ומסודר.

איור 2. האוואטאר מרקוס הפורמלי.

מרקוס הספורטיבי, המופיע באיור 3, עוצב כלובש חליפת ספורט ירוקה, כובע צמר וזקן קצר.

איור 3. האוואטאר מרקוס הספורטיבי.

מרקוס במראה היום-יומי, המופיע באיור 4, עוצב כלובש ג׳ינס וחולצה קצרה צבעונית ובעל תספורת מתולתלת.

איור 4. האוואטאר מרקוס במראה היום-יומי.

מהלך המחקר

הניסוי נחלק לשלושה חלקים. בשלב הראשון קיבלו המשתתפים הסבר על הניסוי והדגמה לצורת השימוש במערכת בנוסף לדף עזר לשימוש במערכת בעת הניסוי. לאחר ההסבר, ענו המשתתפים על שאלון שבו מילאו פרטים על עצמם (גיל, מגדר, מחלקת לימוד במכללה ועוד).

בחלק השני ביצעו המשתתפים את משימתם במערכת EON XR – אימון באמצעות שני ראיונות עבודה עם האוואטאר שאליו שובצו. כל ראיון נחלק לשלבים הבאים: למידה, אימון והערכת ביצוע. שלב הלמידה היה זהה לכל הקבוצות והועבר על ידי האוואטאר המנחה ארתור. האוואטאר הסביר נושאים חשובים לקראת ביצוע ראיונות עבודה, העלה עקרונות מהותיים ואפשר למשתתפים לצפות בסרטון מתאים לכל נושא שהוסבר. ניתן לראות זאת באיור 5.

איור 5. שלב הלימוד – האוואטאר ארתור בעת ביצוע הסבריו על הנושאים השונים בראיונות עבודה.

לאחר מכן, בשלב האימון, המשתתפים ביצוע את ראיון העבודה עם אחד מהאוואטארים בשם מרקוס שאותו קיבלו כמראיין. בראיון שאל מרקוס את המשתתפים שאלות אשר מצופה שיעלו בעת ראיון עבודה אמיתי (כגון ״ספר/י לי על עצמך, ״ספר/י לי על מקרה שבו הצלחת להתמודד עם מצב לחץ״, ״האם את/ה עובד/ת טוב בתנאי לחץ?״, ועוד). התקשורת עם האוואטאר הייתה באמצעות דיבור למיקרופון של אוזניות ששימשו גם על מנת

לשמוע את האוואטאר ואת השאלות ששאל באמצעות דיבור בקול אנושי. המשתתפים הקליטו את תשובותיהם על ידי לחיצה על כפתור מתאים במסך, הקלט הומר לפרומפט שקלט האוואטאר, וזה השיב בהתאם לקלט שקיבל. כל השיחה, השאלות והתשובות של האוואטאר ושל המשתתפים הופיעו בצד המסך בצ׳אט, כאשר תשובות המשתתפים הופיעו בצבע כחול ושאלות האוואטאר באפור. זאת ניתן לראות באיור 6.

איור 6. שלב האימון – צ׳אט התקשורת עם האוואטאר בעת סימולציית ראיון העבודה.

כאשר האוואטאר קיבל מספיק מידע הוא הפסיק את הראיון על ידי הודעה מתאימה כמו ״יום טוב!״. כך ידעו המשתתפים שהראיון נגמר והם עברו לשלב הבא, הערכת הביצועים. היות ומדובר ב-ChatGPT שמבצע את הראיון, הראיונות לא היו זהים לגמרי, והשלב הבא היה תלוי בכך ולכן גם הוא היה לא זהה בין כולם.

בשלב הערכת הביצועים, הציג האוואטאר ארתור מהשלב הראשון דיאגרמת הערכת ביצועים בנושאים ויכולות שונות, שבה הוצג כמה הצליחו להראות במהלך הראיון יכולות בשלל נושאים (לדוגמה מקצועיות, עבודה תחת לחץ, יחסי אנוש ועוד). מעל לדיאגרמה הופיע ציון משוקלל שקיבלו המשתתפים עבור הראיון שחושב על פי הציון שקיבלו לכל קטגוריה שהוצגה ושאותה בחן האוואטאר במהלך הראיון. בנוסף לדיאגרמה הוצג למשתתפים גם סיכום של הראיון בו הוצגו עיקרי הדברים מהראיון. לאחר מכן הוצג בפניהם גם משוב מפורט על מהלך הראיון ונקודות והצעות לשיפור לראיונות הבאים. כל אלו אפשרו למשתתפים להבין היכן כדאי להם להשתפר לקראת הראיונות הבאים. דוגמה לדיאגרמת היכולות מופיעה באיור 7.

איור 7. שלב הערכת הביצועים – דוגמה לדיאגרמת יכולות בה נראה ציון סופי של 28% לראיון העבודה.

לאחר שלב הערכת הביצוע, המשיכו המשתתפים לבצע את כל שלושת השלבים מחדש עם אותו אוואטאר מראיין, כך שיכלו לנסות ולהשתפר מפעם שעברה לפי המשוב שקיבלו. גם לאחר הראיון השני קיבלו המשתתפים את המשוב שלהם ואת הערכת הביצוע שלהם בראיון השני.

לאחר ביצוע שני הראיונות עברו המשתתפים לחלק האחרון בניסוי, בו מילאו שאלון שביעות רצון לגבי העזרה של המערכת בשיפור ביצועיהם והאינטראקציה איתה. כל שאלה דורגה באמצעות סולם ליקרט, מספר בין 1 ל-5, כאשר 1 – לא מסכים, ו-5 – מסכים מאוד.

תוצאות

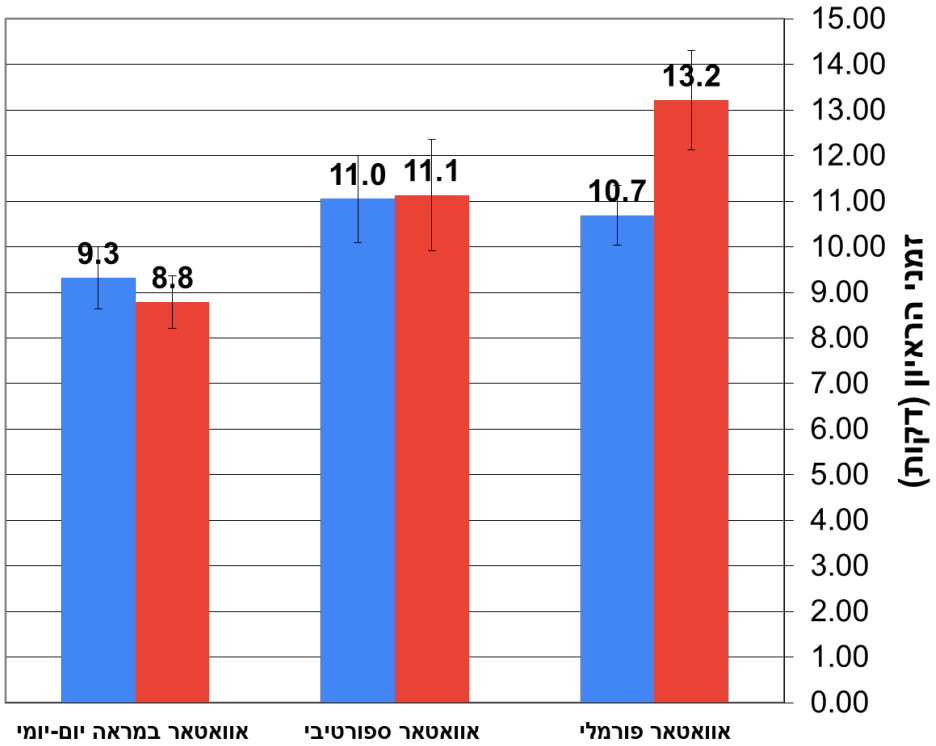
אמצעי מדידה וחישוב

ביצענו Repeated–Measures ANOVAs עבור זמני האימון (הזמן שארך כל ראיון) ועבור הציונים הסופיים (שניתנו לכל המשתתפים בסוף כל ראיון), עם הקבוצה שאליה שויכו (אוואטאר פורמלי, ספורטיבי או במראה יום-יומי) והמגדר (נשים או גברים) כמשתנים הבלתי תלוי בין הקבוצות, ומספר הראיון (ראיון ראשון או שני) כמשתנה הבלתי תלוי בתוך הקבוצות.

בנוסף, ביצענו Univariate ANOVAs עבור כל שאלות שביעות הרצון משאלון הסיום, עם הקבוצה של האוואטאר כמשתנה הבלתי תלוי.

זמני האימון

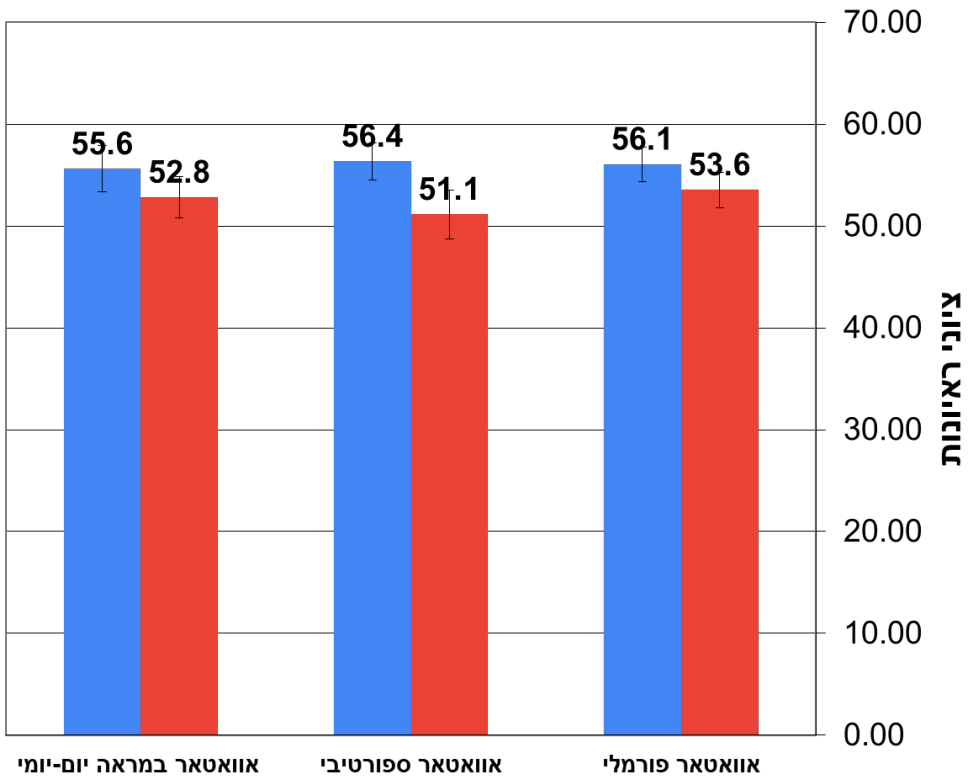
זמן האימון לא היה שונה באופן משמעותי בין שני הראיונות שעברו המשתתפים (עבור הראיון הראשון: $M = 11.1$ דקות, $SD = 5.6$; עבור הראיון השני: $M = 11.4$ דקות, $SD = 4.2$; $F(1,78) = 0.45$, $p = 0.50$). ההבדל בין הקבוצות היה מובהק ($F(2,78) = 5.74$, $p = 0.005$). לכן, ביצענו מבחן LSD (least significant difference) post hoc, אשר הראה שזמן האימון היה קצר יותר לקבוצת האוואטאר במראה יום-יומי ($M = 9.1$ דקות, $SD = 3.3$) בהשוואה לקבוצת האוואטאר הפורמלי ($M = 12.0$ דקות, $SD = 5.0$; $p = 0.005$) ולקבוצת האוואטאר הספורטיבי ($M = 11.1$ דקות, $SD = 5.8$, $p = 0.047$). ההבדל בין זמני האימון של קבוצת האוואטאר הפורמלי וקבוצת האוואטאר הספורטיבי לא היה מובהק ($p = 0.39$). האינטראקציה בין מספר הראיון (ראשון או שני) והקבוצות לא הייתה מובהקת ($F(2,78) = 2.74$, $p = 0.07$). כמו כן לא היה הבדק מובהק בין זמן האימון של נשים ($M = 9.9$ דקות, $SD = 4.5$) וגברים ($M = 10.7$ דקות, $SD = 4.0$; $F(2,78) = 2.31$, $p = 0.11$) והאינטראקציה בין מגדר לקבוצה לא הייתה מובהקת ($F(2,78) = 2.23$, $p = 0.11$). תוצאות אלה מוצגות באיור 8.



איור 8. גרף המציג את ממוצע זמני הראיונות הראשון (באדום) והשני (כחול) לכל קבוצה, עם קווי שגיאות סטנדרטיים.

ציוני ראיונות

הציון הסופי בראיון השני היה גבוה יותר באופן מובהק ($M = 56.0$, $SD = 10.2$) בהשוואה לראיון הראשון ($M = 52.5$, $SD = 10.8$; $F(1,78) = 10.40$, $p = 0.002$). עם זאת, הציונים הסופיים לא היו שונים באופן מובהק בין הקבוצות (עבור קבוצת האוואטאר הפורמלי: $M = 54.8$, $SD = 9.2$; עבור קבוצת האוואטאר הספורטיבי: $M = 54.2$, $SD = 11.4$; עבור קבוצת האוואטאר במראה היום-יומי: $M = 53.8$, $SD = 11.4$; $F(2,78) = 0.07$, $p = 0.94$). האינטראקציה בין מספר הראיון והקבוצות לא הייתה מובהקת ($F(2,78) = 0.16$, $p = 0.85$). כמו כן לא היה הבדק מובהק בין הציון הסופי של נשים ($M = 55.6$, $SD = 11.3$) וגברים ($M = 56.2$, $SD = 9.4$; $F(2,78) = 0.69$, $p = 0.50$) והאינטראקציה בין מגדר לקבוצה לא הייתה מובהקת ($F(2,78) = 0.45$, $p = 0.64$). תוצאות אלו ניתן לראות באיור 9.



איור 9. גרף המציג את ממוצע ציוני הראיונות שקיבלה בראיון הראשון (בכחול) והשני (באדום) כל קבוצה, עם קווי שגיאות סטנדרטיים.

שאלות שביעות רצון

בכל חמש השאלות הנוגעות לשביעות רצונם של המשתתפים, ההבדל בין הקבוצות לא היה מובהק. ראו בטבלה מס׳ 1.

טבלה 1. נתונים אודות דירוגי שאלות שביעות הרצון בכל קבוצה

שאלה	קבוצת האוואטאר הפורמלי	קבוצת האוואטאר הספורטיבי	קבוצת האוואטאר במראה היום-יומי	F	p
"האוואטאר עזר לי להשתפר בראיון העבודה"	$M = 3.4$, $SD = 0.9$	$M = 3.5$, $SD = 0.9$	$M = 3.6$, $SD = 1.0$	$F(2,82) = 0.57$	$p = 0.57$
"אני סומך/ת על ההמלצות שקיבלתי מהאוואטאר"	$M = 3.6$, $SD = 1.2$	$M = 3.8$, $SD = 0.6$	$M = 3.7$, $SD = 0.8$	$F(2,82) = 0.37$	$p = 0.70$
"הרגשתי טוב במהלך הניסוי"	$M = 3.6$, $SD = 1.0$	$M = 3.6$, $SD = 1.1$	$M = 3.5$, $SD = 1.0$	$F(2,82) = 0.12$	$p = 0.89$
"המשוב מהאוואטאר היה מועיל"	$M = 3.8$, $SD = 1.2$	$M = 3.9$, $SD = 0.9$	$M = 3.8$, $SD = 0.8$	$F(2,82) = 0.18$	$p = 0.83$
"המשוב מהאוואטאר היה ברור"	$M = 4.0$, $SD = 1.0$	$M = 4.2$, $SD = 0.8$	$M = 4.3$, $SD = 0.8$	$F(2,82) = 0.61$	$p = 0.54$

דיון ומסקנות

השערותינו התקבלו באופן חלקי. ההבדל בין הקבוצות ניכר רק בזמני האימון, ורק קבוצת האוואטאר במראה היום-יומי בלטה בכך. זמן האימון של קבוצת האוואטאר במראה היום-יומי היה קצר יותר מזה של קבוצות האוואטארים הספורטיבי והפורמלי. המערכת סיימה את האימון כאשר נענו כל שאלות האוואטאר, ולכן זמן האימון הקצר עשוי להצביע על תשובות קצרות יותר של המשתתף. ייתכן כי הדבר נבע מאמון נמוך יותר של המשתתפים במערכת ורמת שיתוף פעולה נמוכה יותר במשחק התפקידים כשהאוואטאר היה בעל מראה שהוא מאוד לא רשמי.

באשר לציונים הסופיים, כל המשתתפים בכל הקבוצות השתפרו בראיון השני לעומת הראשון, אך לא היו הבדלים בין הקבוצות בציונים או ברמת השיפור. ייתכן שהדבר מעיד על כך שמראה האוואטאר לא השפיע על הביצועים בפועל במהלך הראיון. עם זאת, מאחר שזמן האימון של קבוצת האוואטאר במראה היום-יומי היה קצר יותר, ייתכן שזה נובע מחוסר רגישות בהערכת הביצועים שמבצעת המערכת. הסבר נוסף הוא שהתשובות של קבוצת האוואטאר במראה היום-יומי היו קצרות אך באותה איכות כמו של שתי הקבוצות האחרות. מחקר עתידי יכול לחקור לעומק את הסיבות לממצאים אלו. גם מן השאלות בנוגע לשביעות רצונם של המשתתפים לא עלה הבדל משמעותי בין הקבוצות. בנוסף, הדירוג היה דומה כמעט לכל השאלות, בין 3 ("מסכים חלקית") ל-4 ("מסכים").

לסיכום, מערכת האימון לראיון עבודה עם אוואטאר מבוסס ChatGPT אפשרה למשתתפים לשפר את כישוריהם, כפי שנראה בציונים הגבוהים יותר בראיון השני, והמשתתפים תפסו אותה באופן חיובי. מראה האוואטארים השפיע על זמן האימון, שהיה קצר יותר עבור משתתפים שפגשו את האוואטאר בעל המראה היום-יומי. יש לציין שייתכן ומספר הנבדקים הקטן מנע הסקת מסקנות משמעותיות יותר, ואנו מציעים שבמחקרי המשך מספר הנבדקים יהיה גדול יותר. אף על פי שהממצאים אינם חד-משמעיים, ייתכן שיש להם השלכות מעשיות בתכנון צ'אטבוט לאימון לקראת ראיונות עבודה מבוסס LLM. מומלץ שאינטראקציה כזו עם המשתתף תתבצע עם אוואטאר בעל מראה רשמי, או לפחות לא במראה יום-יומי. בנוסף, לגבי ההשלכות על ביצוע אימון לשיפור כישורים רכים אחרים באמצעות משחקי תפקידים, מומלץ שמראה האוואטאר ישקף נורמות חברתיות אמיתיות המתאימות לסיטואציה שאותה פוגשים המשתמשים.

עולה השאלה האם המסקנות וההמלצות ממחקרנו חלות רק על מערכות אימון מבוססות LLM, או שניתן להכליל אותן למערכות צ'אטבוט אחרות, לא בהכרח מבוססות LLM. למערכות מבוססות LLM יש תכונות ייחודיות שמבדילות אותן ממערכות אימון קונבנציונליות. הן מאפשרות למשתתפים לתקשר איתן בשפה טבעית והן מסוגלות לחקות התנהגות אנושית במידה רבה. כתוצאה מכך, יש להן פוטנציאל לעודד משתתפים לשתף פעולה במשחקי תפקידים לצורך אימון או הכשרה במגוון תחומים. לפיכך, ההמלצה לעצב את מראה האוואטאר כך שיתאים לדמות אופיינית במשחק התפקידים עשויה להיות חשובה יותר עבור מערכת מבוססת LLM. מחקר עתידי יכול לחקור נושא זה עם מערכות אימון שאינן מבוססות LLM.

תודות

המחקר מומן באופן חלקי על ידי המועצה להשכלה גבוהה, ישראל.

מקורות

Claude.AI, 2024. https://claude.ai/login?returnTo=%2F%3F.

Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human-Computer Interaction, 19(1)*, 151–181. https://doi.org/10.1207/s15327051hci1901&2_7.

Huang, J., Zhou, M., & Yang, D. (2007). Extracting chatbot knowledge from online discussion forums. *Proceedings of the 20th IJCAI – International Joint Conferences on Artificial Intelligence* (Vol. 7, pp. 423–428). https://doi.org/10.1016/j.compedu.2022.104646.

Lee, K. M., Park, N., & Song, H. (2005). Can a robot be perceived as a developing creature? Effects of a robot's long-term cognitive developments on its social presence and people's social responses toward it. *Human Communication Research, 31(4)*, 538–563. https://doi.org/10.1111/j.1468-2958.2005.tb00882.x.

Liew, T. W., Tan, S. M., Tee, J., & Goh, G. G. G. (2021). The effects of designing conversational commerce chatbots with expertise cues. *2021 14th International Conference on Human System Interaction (HSI)*, 1–6. https://doi.org/10.1109/HSI52170.2021.9538741.

Llanes-Jurado, J., Gómez-Zaragozá, L., Minissi, M. E., Alcañiz, M., & Marín-Morales, J. (2024). Developing conversational Virtual Humans for social emotion elicitation based on large language models. *Expert Systems with Applications*, *246*, 123261. https://doi.org/10.1016/j.eswa.2024.123261.

Microsoft, 2024. https://copilot.microsoft.com/

OpenAI, 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt.

Parise, S., Kiesler, S., Sproull, L., & Waters, K. (1999). Cooperating with life-like interface agents. *Computers in Human Behavior, 15(2)*, 123–142. https://doi.org/10.1016/S0747-5632(98)00035-1

Reeves, B., & Nass, C. (1996). *The media equatin: How people treat computers, television, and new media like real people*. Cambridge university Press.

Riggio, R. E. (1986). Assessment of basic social skills. *Journal of Personality and Social Psychology*, *51*(3), 649. https://doi.org/10.1037/0022-3514.51.3.649

Sonlu, S., Bendiksen, B., Durupinar, F., & Güdükbay, U. (2024). The effects of embodiment and personality Expression on learning in LLM-based educational agents. *arXiv preprint arXiv:2407.10993*. https://doi.org/10.48550/arXiv.2407.10993.

Wan, H., Zhang, J., Suria, A. A., Yao, B., Wang, D., Coady, Y., & Prpa, M. (2024). Building LLM-based AI agents in social virtual reality. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–-7). https://doi.org/10.1145/3613905.3651026.

Wang, Y., Song, M., Guo, R., & Duan, Y. (2023). How about non-human tour guides? The influence of AI tour guides' dress and conversation style on the intention of consumers to continue using them. *Journal of Travel & Tourism Marketing*, *40*(9), 849–862. https://doi.org/10.1080/10548408.2023.2296638.

Yamazaki, T., Mizumoto, T., Yoshikawa, K., Ohagi, M., Kawamoto, T., & Sato, T. (2023). An open-domain avatar chatbot by exploiting a large language model. *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 428–432). cdoi.org/10.1016/j.psychres.2021.114135.

Yang, D., Ziems, C., Held, W., Shaikh, O., Bernstein, M. S., & Mitchell, J. (2024). Social skill training with large language models. *arXiv preprint arXiv:2404.04204*. https://doi.org/10.48550/arXiv.2404.04204.