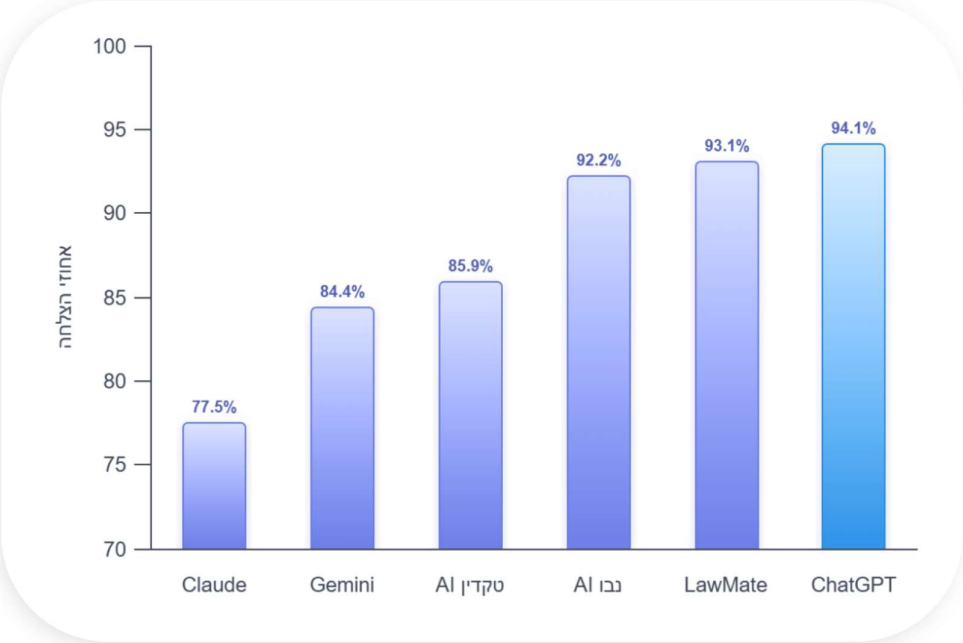
כלי בינה מלאכותית עוברים את בחינות לשכת עורכי הדין בישראל בהצלחה מסחררת – אבל זה לא כל הסיפור

LizzyAI Ltd.

מאמר זה סוקר את תוצאותיה של בדיקת יכולתם של כלי בינה מלאכותית לפתור בהצלחה את בחינות לשכת עורכי הדין בישראל, תוך התמקדות בחלק השאלון האמריקאי במבחן. במסגרת הבדיקה נבחנו שישה כלים – ChatGPT, Claude, Gemini, נבו AI, טקדין AI ו-LawMate – בשתי מתודולוגיות: פתרון כל אחד מהחלקים ברצף, לצד פתרון של כל שאלה בנפרד.

הממצאים מלמדים על רמת ביצועים גבוהה של הכלים. כלל הכלים השיגו תוצאות מספקות, ומרביתם אף הגיעו לשיעורי הצלחה גבוהים במיוחד. הממצאים מעלים תובנות חשובות בנוגע להמשך הטמעתם של כלי בינה מלאכותית בעשייה המשפטית, הנוגעות לא רק ליכולת הכלים להשיב תשובה נכונה באופן פשטני, אלא ביחס ליכולות ולמאפיינים הנדרשים מהם כדי להתאים לצרכי התחום ולתרום לו באופן ממשי.

<u>תוצאות מועד בחינות הלשכה האחרון – יוני 2026</u>



1. מבוא

אנחנו מצויים בעיצומה של מהפכה חסרת תקדים. השימוש בכלי בינה מלאכותית הפך לחלק אינטגרלי משוק העבודה, הן במשימות יומיומיות והן בתהליכים מורכבים, וכלי בינה מלאכותית מפגינים יכולות מתקדמות יותר ויותר בתחומים הדורשים חשיבה לוגית ויכולות אינטגרטיביות ואנליטיות. מהפכה זו לא פסחה על עולם המשפט, וכיום קיימים מספר כלי בינה מלאכותית המסוגלים לבצע ברמה גבוהה משימות הנדרשות במסגרת מתן שירותים משפטיים, מחקר משפטי ועוד. על רקע זה, החלטנו לבדוק האם כלים אלו מסוגלים כבר היום לפתור בהצלחה את בחינות לשכת עורכי הדין בישראל.

בדיקה זו נולדה מצורך מעשי: במסגרת הליכי הפיתוח של מאפיין (Feature) חדש של מאגר משפטי ב-LizzyAI, ביקשנו לבחון חלופות שונות לקראת ההשקה ולהתרשם מיכולותיהם של הכלים הזמינים להשיב על שאלות משפטיות בצורה מדויקת ומהימנה. כמו כן, ראינו בכך הזדמנות לתרום נדבך נוסף לשיח המקצועי והאקדמי בנושא השפעת הבינה המלאכותית על עולם המשפט.

שאלת יכולתם של כלי בינה מלאכותית לעבור בחינות הסמכה משפטיות עוררה עניין גם במדינות נוספות. לדוגמה, במחקר שפורסם בשנת 2024 בכתב העת Philosophical Transactions of the Royal Society A, נבחנו ביצועיו של GPT-4 במבחן ההסמכה האמריקאי לעריכת דין (UBE), ונמצא כי המודל עבר את המבחן ואף השיג תוצאה העולה על ממוצע הניגשים לבחינה.[1]

גם בישראל נבחנה שאלה דומה. בדצמבר 2025 בחן רז ברכה את ביצועיהם של Gemini, ChatGPT, נבו AI וטקדין AI במועד דצמבר 2025 של בחינות הלשכה.[2] הבדיקה שלנו מרחיבה ומעמיקה את הבחינה הקודמת במספר היבטים: מספר מועדי הבחינה ומספר הכלים שנבדקו, מתודולוגיות הבדיקה, תחקור טעויות הכלים וגיבוש התובנות מהממצאים.

2. מתודולוגיה

בחינת לשכת עורכי הדין בישראל מורכבת משלושה חלקים: מטלת כתיבה, חלק דיוני וחלק מהותי. במטלת הכתיבה, שמשקלה 20% מהציון, מקבלים הנבחנים תיאור מקרה על בסיסו הם נדרשים לנסח את המסמך המשפטי המתאים להגשה לבית המשפט. החלק הדיוני והחלק המהותי, אשר משקל כל אחד מהם הוא 40% מהציון, בנויים כל אחד מ-40 שאלות אמריקאיות (רב-ברירה) עם ארבע תשובות אפשריות. החלק הדיוני מתמקד בפן הפרוצדורלי של המשפט, בעוד שהחלק המהותי נוגע בסוגיות הדין המהותי בתחומי המשפט האזרחי, הפלילי והמנהלי. הציון הכולל הנדרש למעבר המבחן הוא 60 נקודות.

<u>היקף הבדיקה</u>

הבדיקה התמקדה בחלק הדיוני ובחלק המהותי בלבד, ולא כללה את חלק הניסוח המשפטי (מטלת הכתיבה). הבדיקה נערכה על פני ארבעה מועדי בחינות ובשתי מתודולוגיות: במועד דצמבר 2025 נבדקו כלל הכלים הן בהרצת כל חלק בבת אחת והן בהרצת כל שאלה בנפרד בשיחה חדשה. במועדי יוני 2025 ודצמבר 2024 נבדקו שלושת הכלים המובילים בבדיקת מועד דצמבר 2025 בלבד, במתודולוגיה של הרצת כל חלק בבת אחת.

---

[1] Daniel Martin Katz, Michael James Bommarito, Shang Gao, Pablo Arredondo *GPT-4 passes the bar exam*, 382 PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A 1 (2024).

[2] קישור לפוסט של רז ברכה – LinkedIn.

הבדיקה נערכה ברובה בחודשים מרץ ואפריל 2026, ונעשה בה שימוש בשישה כלי בינה מלאכותית, אותם חילקנו לשתי קטגוריות. קטגוריית הכלים ה"בסיסיים" כללה את שלושת מודלי שפה הגדולים ChatGPT, Claude ו-Gemini, וקטגוריית הכלים ה"משפטיים" כללה שלושה כלים ישראליים מבוססי LLM שמיועדים לעבודה משפטית – נבו AI, טקדין AI ו-LawMate. בכלים המאפשרים זאת, השתמשנו במודלים ובמצבי החשיבה כפי שיפורטו להלן. לקראת סיום כתיבת מאמר זה פורסמו התשובות הרשמיות של מועד יוני 2026 מטעם לשכת עורכי הדין, והוחלט לכלול גם אותו במסגרת הבדיקה, אולם בשימוש הגרסאות העדכניות של הכלים לאותה העת. על מועד יוני 2026 נערכו בדיקות במתודולוגיה של כל חלק בבת אחת, ובמתודולוגיה של כל שאלה בנפרד. במסגרת בדיקת כל שאלה בנפרד בוצעו ארבע הרצות, והתוצאות המפורטות במאמר זה מהוות את ממוצע ארבע ההרצות.[3]

1. **הכלים ה"בסיסיים":***
   (1) **ChatGPT** 5.4 – Extended Thinking
   (2) **Claude** Opus 4.6 – Extended Thinking
   (3) **Gemini** 3.1 – Pro

*בבדיקת מועד יוני 2026, נעשה שימוש בגרסאות GPT-5.6 Sol (High Intelligence), Opus 4.8 (High Effort, Thinking) ו-3.6/Gemini Flash 3.5 (Thinking).[4]

2. **הכלים ה"משפטיים":**
   (4) **נבו AI**
   (5) **טקדין AI**
   (6) **LawMate**

<u>אופן הבדיקה</u>

הן בהרצת כל חלק בבת אחת והן בהרצת כל שאלה בנפרד, כל שאילתה הוזנה בשיחה חדשה. בהרצות של כל חלק בבת אחת, העתקנו את השאלות משאלון המבחן הרשמי והדבקנו אותן כטקסט בשדה הקלט של כל כלי.[5] השתמשנו ב-Skill שיצרנו ב-Lizzy כדי להמיר בצורה מהירה את התשובות שסיפק כל כלי לעמודת אותיות שהוזנה למאגר המעקב שיצרנו.

בהרצות של כל שאלה בנפרד, העתקנו את השאלה והדבקנו אותה כטקסט בשדה הקלט, בתוספת בקשה לציין את האות שמסמלת את ההיגד הנכון מבין ארבעת ההיגדים (ארבע התשובות). מעבר לבדיקת נכונות המענה עצמו, בוצע גם ניתוח של תוכן המענה שסיפקו הכלים, במטרה לגבש תמונה מלאה על יכולתו של הכלי לסייע בעבודה משפטית ולזהות דפוסי טעויות אפשריים.

<u>מהימנות הבדיקה</u>

לצורך הבטחת מהימנות ואחידות הבדיקה, בבדיקת כל חלק בבת אחת בוצעו חמש הרצות לכל כלי, והנתונים הכלולים בממצאים משקפים את ממוצע ההרצות. בבדיקת כל שאלה בנפרד, תועדה בממצאים התשובה הראשונה שסיפק כל כלי.

---

[3] מידת השונות בין ארבע ההרצות הייתה נמוכה יחסית.

[4] ב-Gemini נבחנו הן מודל 3.1 Pro והן מודל 3.5 Flash ו-3.6 Flash. מודלי Flash השיגו תוצאות גבוהות יותר, וייתכן שהדבר נובע, בין היתר, מכך שהם חדש יותר. לפיכך, החלטנו להציג את תוצאותיהם, אף שאינם המודלים החזק ביותר של Gemini.

[5] זאת למעט ב-LawMate, בו בעת ביצוע הבדיקה הייתה מגבלה על מספר התווים שניתן להדביק בשדה הקלט, כך שהשאלות צורפו כקובץ PDF בתוך מצב "ניתוח מסמכים מוכוון מחקר משפטי"; ולמעט בטקדין AI, אשר גם בו, בעת ביצוע הבדיקה, הייתה מגבלה על מספר התווים שניתן להדביק, ואף לא היה ניתן לצרף את השאלות כקובץ מצורף. מסיבות אלו ההרצה של כל חלק בבת אחת לא נבחנה בטקדין AI.

מעבר לבדיקות העיקריות שמטרתן לבחון את נכונות התשובה, בהרצת כל שאלה בנפרד נערכו בדיקות נוספות כדי לבחון את איכות המענה של הכלים מבחינת עקביות התשובות. בכל אחד מהכלים נבחרו באופן מדגמי שאלות שנשאלו מספר פעמים, כדי לבדוק האם הכלי עקבי בתשובותיו. כמו כן, נערכו בדיקות שבהן הוחלף סדר התשובות, כדי לבחון את יכולת הניתוח המהותי של החלופות ולוודא שהמענה אינו נשען רק על זיהוי טכני של תבניות.

<u>מגבלות הבדיקה</u>

בבדיקה שערכנו קיימות מספר מגבלות מובנות. המגבלה הראשונה היא היקף הבדיקה. נבדקו ארבע בחינות בלבד, כאשר בארבעת המועדים נערכו הרצות של כלל החלקים בבת אחת ובמועד דצמבר 2025 נערכו בנוסף הרצות של כל שאלה בנפרד. כמו כן, הבדיקה נערכה על מספר מוגבל של כלים. כמובן שככל שבדיקה תהא מקיפה יותר ותכלול מספר גדול יותר של מועדים וכלים, היא תוכל לספק תמונה רחבה ומהימנה יותר של יכולותיהם.

כמו כן, ישנו סיכון שהכלים ישתמשו בזיכרון שנאסף על העדפות המשתמש או בהפניה לשיחות קודמות. מצב זה עלול לגרום לכך שהכלי יסתמך על תשובותיו הקודמות, וכך המענה ישתפר מפעם לפעם ולא ישקף בצורה מהימנה את יכולתו של הכלי לספק תשובה נכונה לשאלה שלא נתקל בה בעבר. בבדיקות שערכנו בכלים הבסיסיים, בהם הגדרות אלו פתוחות לשינוי עבור המשתמש, הגדרות הזיכרון וההפניות לשיחות קודמות היו מושבתות. כדי לבחון מגבלה זו בכל זאת, ביצענו בדיקה מדגמית ב-ChatGPT שבה השווינו בין שיעורי ההצלחה כאשר הגדרות הזיכרון הופעלו לבין שיעורי ההצלחה במצב בו אותן הגדרות היו מושבתות. אם אכן נעשה שימוש בזיכרון, היינו מצפים לראות שיפור הדרגתי בשיעורי ההצלחה לאורך ההרצות – אך התוצאות שקיבלנו היו זהות, או עם הבדלים זניחים.

מגבלה נוספת היא האפשרות שבמסגרת חיפוש התשובות לשאלות, הכלים יאתרו את מפתח התשובות הרשמי של לשכת עורכי הדין ברשת ויסתמכו עליו. פרקטיקה זו רצויה בשימוש יומיומי, שכן היא מאפשרת לכלי לאמת ולבסס את תשובתו. ואולם, במסגרת בדיקה זו מדובר במגבלה שעלולה להביא לכך שהממצאים לא ישקפו את יכולותיהם האנליטיות האמיתיות של הכלים. חלק מהכלים מציגים למשתמש את המקורות החיצוניים – כלומר, מקורות שנובעים מחיפוש ברשת ולא מחיפוש במאגר פנימי – שעליהם התבססו בחיפוש התשובה. בחלק מההרצות הכלים אכן איתרו את מפתח התשובות הרשמי ונעזרו בו במסגרת המענה. עם זאת, באופן מפתיע, גם כאשר בוצעה הסתייעות כזו, הכלים עדיין לא סיפקו מענה נכון לכל השאלות. על כן, נראה כי מפתח התשובות לא שימש בסיס להעתקה ישירה וטכנית של התשובות, אלא כמקור נוסף לצורך בדיקת המענה, בדומה למקורות אחרים שבהם נעשה שימוש. כמו כן, נערכו בדיקות מדגמיות שכללו בקשה להימנע מעיון במפתח התשובות הרשמי, ונמצא כי הפער בשיעורי ההצלחה בבדיקות אלו היה זניח.

נוסף על המגבלות שפורטו לעיל, קיימת מגבלה הנובעת באופן טבעי מעדכוני הגרסאות של הכלים. כלי בינה מלאכותית מצויים בתהליך שיפור מתמיד, וגרסאותיהם מתעדכנות בקצב מהיר. בדיקה זו משקפת את היכולות של הגרסאות המסוימות של הכלים שהיו זמינות במועד עריכת הבדיקה, כאשר במועד פרסומו של מאמר זה סביר כי לכל אחד מן הכלים כבר קיימות גרסאות חדשות יותר, אשר יציגו יכולות מרשימות ומדויקות יותר מאלו המובאות בממצאי המאמר. על כן, הבדיקה שערכנו משקפת טעימה מיכולותיהם של הכלים והשוואה מעניינת ביניהם ברגע נתון, שמטרתה לספק נקודה למחשבה באשר ליכולות הנדרשות מהבינה המלאכותית לצורך השתלבות בעבודה המשפטית.

3. הממצאים

הממצאים המוצגים משקפים את סך הטעויות ואחוזי ההצלחה של כל כלי בחלק המהותי ובחלק הדיוני יחד, כלומר, ב-80 שאלות בסך הכל. הציונים עוגלו למספר השלם הקרוב, כאשר חצי נקודה ומעלה עוגלו כלפי מעלה.

<u>יוני 2026</u>

<u>הרצת כל שאלה בנפרד</u>

	ChatGPT	Gemini	Claude	LawMate	נבו AI	טקדין AI	ממוצע
מספר טעויות	4.8	12.5	18	5.5	6.3	11.3	9.7
אחוזי הצלחה	94.1%	84.4%	77.5%	93.1%	92.2%	85.9%	87.9%

ChatGPT הגיע למקום הראשון עם 94.1% הצלחה (4.8 טעויות), אחריו **LawMate** עם 93.1% הצלחה (5.5 טעויות) **ונבו AI** עם 92.2% הצלחה (6.3 טעויות). אחריהם דורגו **טקדין AI** עם 85.9% הצלחה (11.3 טעויות), **Gemini** עם 84.4% הצלחה (12.5 טעויות) ו-**Claude** עם 77.5% הצלחה (18 טעויות).

<u>הרצת כל חלק בבת אחת</u>

	ChatGPT	Gemini	Claude	LawMate	נבו AI	ממוצע
מספר טעויות	4.4	14	18.4	17	16	14
אחוזי הצלחה	94.5%	82.5%	77%	79%	81%	82.8%

בבדיקה זו **ChatGPT** מוביל עם 94.5% הצלחה (4.4 טעויות), אחריו **Gemini** עם 82.5% הצלחה (14 טעויות), **נבו AI** עם 81% הצלחה (16 טעויות), **LawMate** עם 79% הצלחה (17 טעויות), ולבסוף **Claude** עם 77% הצלחה (18.4 טעויות).

<u>הטוב מבין שתי המתודולוגיות</u>

	ChatGPT	Gemini	Claude	LawMate	נבו AI	טקדין AI	ממוצע
אחוזי הצלחה	96.3%	86.3%	83.8%	95%	95%	88.8%	89.6%

ChatGPT הגיע למקום הראשון עם 96.3% הצלחה, אחריו **LawMate ונבו AI** עם 95% הצלחה כל אחד, לאחר מכן **טקדין AI** עם 88.8% הצלחה ו-**Gemini** עם 86.3% הצלחה. **Claude** מדורג אחרון עם 83.8% הצלחה.

<u>דצמבר 2025</u>

<u>הרצת כל שאלה בנפרד</u>

	ChatGPT	Gemini	Claude	LawMate	נבו AI	טקדין AI	ממוצע
מספר טעויות	6	10	27	4	8	13	11.3
אחוזי הצלחה	92.5%	87.5%	66.3%	95%	90%	83.8%	85.8%

שלושת הכלים המובילים היו **LawMate**, אשר השיג את שיעור ההצלחה הגבוה ביותר עם 95% הצלחה (4 טעויות), **ChatGPT** עם 92.5% הצלחה (6 טעויות), **ונבו AI**, שהגיע ל-90% הצלחה (8 טעויות). אחריהם דורגו **Gemini** עם 87.5% הצלחה (10 טעויות), **טקדין AI** עם 83.8% הצלחה (13 טעויות), ו-**Claude**, שהשיג את התוצאה הנמוכה ביותר עם 66.3% הצלחה (27 טעויות).

הרצת כל חלק בבת אחת

	ChatGPT	Gemini	Claude	LawMate	נבו AI	ממוצע
מספר טעויות	8.3	17.2	21.8	27.4	28	20.5
אחוזי הצלחה	89.6%	78.5%	72.8%	65.8%	65%	74.3%

בהרצת כל חלק בבת אחת הממצאים שונים מאלו שהתקבלו בהרצת כל שאלה בנפרד. בבדיקה זו הכלים הבסיסיים הובילו, כאשר **ChatGPT** הגיע ל-89.6% הצלחה (8.3 טעויות), אחריו **Gemini** עם 78.5% הצלחה (17.2 טעויות), ו-**Claude** עם 72.8% הצלחה (21.8 טעויות). **LawMate** השיג 65.8% הצלחה (27.4 טעויות), ואילו **נבו AI** השיג את התוצאה הנמוכה ביותר עם 65% הצלחה (28 טעויות).

יוני 2025

הרצת כל חלק בבת אחת

	ChatGPT	LawMate	נבו AI	ממוצע
מספר טעויות	8.2	25.2	32.2	21.9
אחוזי הצלחה	89.8%	68.5%	59.8%	72.7%

כאמור לעיל, בבדיקה זו ובבדיקה הבאה נבחנו רק שלושת הכלים שהובילו בבדיקת כל שאלה בנפרד במועד דצמבר 2025. ניתן לראות כי **ChatGPT** מוביל גם בבדיקה זו עם 89.8% הצלחה (8.2 טעויות), וכי קיים פער גדול בינו לבין **LawMate** ו**נבו AI** – **LawMate** הגיע ל-68.5% הצלחה (25.2 טעויות), ו**נבו AI** ל-59.8% הצלחה בלבד (32.2 טעויות).

דצמבר 2024

הרצת כל חלק בבת אחת

	ChatGPT	LawMate	נבו AI	ממוצע
מספר טעויות	14.4	25.6	32.4	24.1
אחוזי הצלחה	82%	68%	59.5%	69.8%

בהשוואה למועד יוני 2025, גם בבדיקה זו **ChatGPT** מוביל, אולם עם שיעורי הצלחה נמוכים יותר – 82% הצלחה (14.4 טעויות). **LawMate** השיג שיעור הצלחה דומה – 68% הצלחה (25.6 טעויות), ואילו **נבו AI** עם שיעור הצלחה דומה של 59.5% (32.4 טעויות).

ממוצע כלל המועדים

הרצת כל חלק בבת אחת

	ChatGPT	Gemini	Claude	LawMate	נבו AI	ממוצע
מספר טעויות	8.8	8.1	20.1	23.8	27.2	17.6
אחוזי הצלחה	89%	80.5%	74.9%	70.9%	66.3%	76.3%

בבדיקת ממוצע כלל המועדים בהרצת כל חלק בבת אחת **ChatGPT** מוביל עם 89% הצלחה (8.8 טעויות), אחריו **Gemini** עם 80.5% הצלחה (8.1 טעויות), **Claude** עם 74.9% הצלחה (20.1 טעויות), **LawMate** עם 70.9% הצלחה (23.8 טעויות), **ונבו AI** עם 66.3% הצלחה (27.2 טעויות).

הרצת כל שאלה בנפרד

	ChatGPT	Gemini	Claude	LawMate	נבו AI	טקדין AI	ממוצע
מספר טעויות	5.4	11.3	22.5	4.8	7.1	12.1	10.5
אחוזי הצלחה	93.3%	85.9%	71.9%	94.1%	91.1%	84.8%	86.8%

בבדיקת ממוצע כלל המועדים בהרצת כל שאלה בנפרד **LawMate** מוביל עם 94.1% הצלחה (4.8 טעויות), אחריו **ChatGPT** עם 93.3% הצלחה (5.4 טעויות), **נבו AI** עם 91.1% הצלחה (7.1 טעויות), **Gemini** עם 85.9% הצלחה (11.3 טעויות), **טקדין AI** עם 84.8% הצלחה (12.1 טעויות), ו-**Claude** עם 71.9% הצלחה (22.5 טעויות).

בדיקות נוספות

בדיקת אקראיות: הכלים המשפטיים התמידו לרוב בעקביות התשובות, ונצפו מקרים ספורים של מתן תשובה שונה לאותה שאלה בהרצות נפרדות. בכלים הבסיסיים, נראו מספר מקרים של מתן תשובות שונות לאותה שאלה בכל הכלים, כאשר מדובר בתדירות גבוהה יותר ב-**Claude** מאשר ב-**Gemini**, ובמקרים בודדים בלבד ב-**ChatGPT**. בבדיקה שבה שונו מיקומי התשובות בשאלה, כל הכלים השיבו באופן עקבי בהתאם לתוכן התשובה, ללא תלות במיקומה.

4. ניתוח הממצאים

אחוזי ההצלחה

מלבד ההבדלים בשיעורי ההצלחה בין הכלים במסגרת כל מועד שנבדק, כפי שפורט בפרק הממצאים, ניתן לזהות הבדל חיובי משמעותי בשיעורי ההצלחה של כל הכלים באופן גורף בהרצת כל חלק בבת אחת בין מועד דצמבר 2025 למועד יוני 2026, אשר סביר כי נובע משיפור ועדכון גרסאות בתקופה שבין הבדיקות.

תוכן התשובות

איכות המענה

מעבר להבדלים בשיעורי ההצלחה במענה, ניתן לזהות בין הכלים פערים גם באיכות המענה מבחינה מהותית, הנוגעים לחוויית המשתמש. פערים אלו באים לידי ביטוי, בין היתר, במהירות המענה, באופן הצגת התשובה מבחינה ארגונית ולוגית, ובעיקר ביכולת המשתמש לאמת את המענה בצורה אפקטיבית מול המקורות שעליהם הוא מבוסס. חוויית האימות היא אחד הפרמטרים המרכזיים בהערכת איכות המענה של כלי, שכן

אין די בקבלת תשובה נכונה לשאלה המשפטית; על מנת להיעזר בכלי בעבודה השוטפת, המשתמש צריך להיות מסוגל לאמת את התשובה בקלות ובמהירות, מבלי להידרש לחיפוש ממושך.

אמנם כלל הכלים מספקים למשתמש את המקורות שעליהם התבססו במתן המענה, אך קיימים הבדלים משמעותיים באופן שבו הם מפנים למקורות אלה. הכלים הבסיסיים מספקים את המקור לתשובתם, אולם לרוב מדובר בהפניה כללית באמצעות קישור חיצוני, ללא מראה מקום מדויק. כמו כן, הכלים הבסיסיים אינם בוחנים בהכרח את מידת המהימנות של המקורות ואינם מסננים אותם בהתאם. לעיתים הכלים הבסיסיים הפנו למקורות שלא נהוג להסתמך עליהם במסגרת מחקר משפטי או מתן שירותים משפטיים, כגון בלוגים ופרסומים פרטיים נוספים ברשת, להבדיל מפסיקה, דברי חקיקה לסוגיהם או ספרות משפטית.

בכלים המשפטיים ההפניה למקורות בוצעה ברמה איכותית יותר, הן מבחינת סוג המקורות והן מבחינת חוויית האימות של המשתמש. ואולם, גם בין כלים אלו קיימים הבדלים באופן ההפניה למקורות: חלק מהכלים מספקים את ההפניה כציטוט מבודד, המלווה בקישור חיצוני למקור המלא. במצב זה, כדי לאמת את המקור ולראות את ההפניה בתוך ההקשר המלא שלו, המשתמש נדרש לעיתים לעבור שלב נוסף של חיפוש ההפניה הספציפית במקור. לעומת זאת, חלק מהכלים מציגים את ההפניה באופן ישיר ובולט בתוך המקור המלא, כך שהמשתמש יכול לבצע אימות מהיר ופשוט. לדעתנו, הדרך האחרונה היא המועדפת.

<u>תחקור טעויות</u>

בשאלות שבהן הכלים השיבו באופן שגוי, מקור הטעות היה בדרך כלל באחד משני מישורים: טעות בדין או בהבנת הכלל המשפטי הרלוונטי, או לחלופין ידיעת הדין אך קושי ביישום נכון שלו על נסיבות המקרה. במקרה האחרון הכלים זיהו את הכלל המשפטי המתאים, אך הניתוח שהציגו היה חלקי ולא כלל התייחסות מספקת לחריגים, סייגים או מקרי קצה רלוונטיים, באופן שגרם לבחירת תשובה שגויה.

<u>השוואה לבדיקות קודמות</u>

הבדיקה שערך עו"ד רז ברכה בינואר 2026 העלתה ממצאים מעט שונים. לפי תוצאות אותה בדיקה, הכלי המוביל היה **נבו AI**, עם 97.5% הצלחה (2 טעויות). אחריו דורגו **Gemini (Pro 3)** עם 83.75% הצלחה (13 טעויות), **ChatGPT (GPT-5.2)** עם 81.25% הצלחה (15 טעויות), ו**טקדין AI** עם 80% הצלחה (16 טעויות). ייתכן כי השוני בין הבדיקות נובע מגרסאות הכלים שבהן נעשה שימוש, מהגדרות הזיכרון של הכלים, מאקראיות טבעית בתשובות וגורמים נוספים.

כמו כן, מעניין להציב את הממצאים מהבדיקה שלנו לצד התוצאות שהתקבלו בבדיקה האמריקאית, שצוינה בתחילת המאמר. גם בבדיקה האמריקאית **ChatGPT** השיג תוצאות מרשימות –**GPT-4** השיג 75.7% הצלחה בחלק שכלל שאלות רב-ברירה (MBE) של מבחן ההסמכה האמריקאי, ציון שעלה על ממוצע הניגשים לבחינה (כ-68% הצלחה). בציון המשוער הכולל, **ChatGPT** צבר כ-297 נקודות – מעל סף המעבר בכל המדינות בארצות הברית שמשתמשות במבחן ה-UBE.

5. מסקנות

בדיקה זו בחנה את יכולותיהם של שישה כלי בינה מלאכותית לעבור בהצלחה את בחינות לשכת עורכי הדין בישראל, על ידי פתרון כל אחד מהחלקים הכוללים שאלות רב-ברירה ברצף, ופתרון של כל שאלה מחלקים אלו בנפרד. הבדיקה המחישה את ההתקדמות המהירה של ביצועי כלי הבינה המלאכותית ככלל, והדגישה את העקרונות המרכזיים לשם מימוש הפוטנציאל בשילוב הבינה המלאכותית במתן שירותים משפטיים בפרט.

מהממצאים עולה כי כלי הבינה המלאכותית מסוגלים להשיג תוצאות גבוהות מאוד בפתרון חלקי רב-הברירה של בחינות הלשכה, ואף להשיג ציון גבוה במידה ניכרת מהציון הנדרש למעבר הבחינה. ChatGPT השיג את התוצאות הגבוהות ביותר במענה על כל חלק בבת אחת, וכן את התוצאות הגבוהות ביותר, לצד LawMate, במענה על כל שאלה בנפרד. ממצאים אלה מעררים את ההנחה לפיה יכולתו של כלי לספק תשובות נכונות נובעת מקיומו של מאגר משפטי פנימי רחב המוטמע בו. הממצאים מעידים על כך שגם יכולות חיפוש איכותי ברשת ויכולות אינטגרציה ולוגיקה גבוהות עשויות להביא למענה נכון בהיעדר מאגר משפטי ייעודי, כפי שממחיש המקרה של ChatGPT.

עם זאת, היכולת לענות נכונה על שאלות משפטיות אינה מספיקה לבדה כדי לאפשר שימוש מעשי בכלי. בעבודה משפטית קיימת חשיבות מכרעת לאיכות ההנמקה, לטיב המקורות וליכולת של המשתמש לאמת את הבסיס המשפטי של המענה. כלי משפטי אידיאלי יהיה כזה שמעבר למתן תשובה נכונה מבחינת הדין, יודע לבצע איזון ושילוב חכם בין מאגר משפטי רחב לבין סוכן חיפוש איכותי, בעל יכולת העמקה וחיפוש ברשת, ומספק למשתמש חוויית אימות מהירה ויעילה למקור המשפטי שעליו מבוססת התשובה.