**דוח מסקנות: דוח בחינת כיול המבקר האוטומטי**

**מוטיבציה לניסוי**

במסגרת הפרויקט, אחת ממטרותינו היא לא רק ליצור סיפורים איכותיים, אלא גם לאפשר להם לעבור תהליך בדיקה אוטומטי המבוסס על עקרונות ביקורת ספרותית מקצועית. ניסוי זה בודק את היכולת של מערכת AI לבקר ולשפר סיפורים ארוכים בעזרת המלצות לתיקון, בהתבסס על תהליך כיול מול ביקורות אנושיות.

**אופן ביצוע הניסוי**

הניסוי התבצע בשני שלבים עיקריים:

1. שלב הכיול:
   * 3 סיפורים קצרים באורך 500 מילים הועברו ביחד עם Prompt למבקר האוטומטי.
   * הציונים וההערות שימשו ללימוד מודל בינה מלאכותית שתפקידה לזהות חולשות מבניות, סגנוניות ועלילתיות בסיפורים.
2. שלב ההשוואה:
   * סיפורים בינוניים באורך כ 2000 מילים הועברו לקריאה והערכה על ידי מבקרים אנושיים, כל סיפור דורג מחדש על ידי המערכת והושווה על בסיס מדדים מוגדרים מראש אל מול הציונים שנתן המבקר האוטומטי.
   * ניתוח ההבדלים נועד לבדוק את תרומת תהליך בדיקת האיכות הכללית של הסיפור.

**כעת נציג את הסבר קצר עבור כל מדד:**

**דמויות:**

במדד זה סוג השאלות שנעזרנו בהן ע"מ לתת דירוג, הן שאלות הבודקות את רקע הדמויות והתפתחותן במהלך העלילה.

**קונפליקט:**

במדד זה סוג השאלות שנעזרנו בהן ע"מ לתת דירוג, הן שאלות הבודקות את העניין בקונפליקט המרכזי והתפתחותו לאורך העלילה ובתתי הקונפליקטים המסתעפים ממנו במהלך העלילה.

**מלאכת הכתיבה:**

במדד זה סוג השאלות שנעזרנו בהן ע"מ לתת דירוג, הן שאלות הבודקות האם הסיפור מנוסח בשפה תקנית ומודרנית.

**היגיון:**

במדד זה סוג השאלות שנעזרנו בהן ע"מ לתת דירוג, הן שאלות הבודקות האם יש היגיון בעלילת הסיפור, למשל האם קיימים חורים או נקודות לא ברורות בעלילה.

**מהלך הניסוי:**

כל סיפור הוערך על פי 30 שאלות ספרותיות מפורטות בארבע קטגוריות מרכזיות:  
 דמויות (Character), קונפליקט (Conflict), סגנון (Craft) והיגיון (Logic) .   
עבור כל קטגוריה הוגדר משקל לאותה הקטגוריה, ועבור כל שאלה בכל אחת מהקטגוריות הוגדר משקל לשאלה על מנת לתת משמעות יחסית לתוצאה הסופית.

שישה סיפורים הוערכו גם על ידי מבקרים אנושיים בצורה ידנית וגם על ידי המבקר האוטומטי. לשם בחינת קצוות, נכללו גם שני מקרים נוספים:

1. סיפור ריק לחלוטין.
2. סיפור מינימליסטי שנכתב בכוונה בצורה גרועה ("The Blue Hat") .

עבור כל אחד מהמדדים השתמשנו בשאלות שהופיעו במאמר להערכת תסריטים תוך התאמת השאלות לסיפורים (ניתן לראות את השאלות בנספח א'), [קישור למאמר](https://www.specscout.com/score).

**תוצאות ההשוואה:**

|  |  |  |
| --- | --- | --- |
| **סיפור** | **ממוצע פער** | **סטיית תקן** |
| Whispers in the Reeds | 0.0083 | 0.0143 |
| The Weight of Legacy | 0.0154 | 0.0236 |
| Ashes of Vengeance | 0.0173 | 0.0202 |
| The Price of Redemption | 0.0218 | 0.0277 |
| Ashes of Vengeance | 0.0262 | 0.0305 |
| Divided Hearts | 0.0222 | 0.0242 |
| (Empty Story) | 0.0000 | 0.0000 |
| The Blue Hat | --- | --- |

**ניתוח ומסקנות:**

לצורך הבנת טיב ההשוואה בין הביקורת האוטומטית לזו האנושית, בוצע ניתוח פרטני לכל אחד מהסיפורים שנבדקו. הניתוח התמקד בזיהוי מגמות חוזרות, אזורי דמיון, פערים בולטים, והסבר לפערים שהתגלו בין הביקורות. להלן עיקרי המגמות עבור כל סיפור בנפרד:

**סיפור 1 - Whispers in the Reeds**

**אזורי דמיון –**

* בפרטי הדמות הראשית והמבנה המרכזי של הקונפליקט הייתה התאמה מוחלטת.
* המבקר האוטומטי זיהה היטב את המורכבות הרגשית והמסלול ההתפתחותי של הגיבור.

**אזורי שונות –**

* לא נמצאו – הפער הממוצע הנמוך ביותר בכל הסיפורים (0.0083).

המבקר האוטומטי העריך את הסיפור ברמה זהה למבקר האנושי כמעט בכל פרמטר.

**סיפור 2 - The Weight of Legacy**

**אזורי דמיון –**

* בשאלות על הדמות הראשית (שאלות 1–6) יש כמעט התאמה מושלמת.
* גם בסעיף הלוגי של קונסיסטנטיות וחוסר בסתירות (שאלות 28–29) הייתה התאמה מלאה.

**אזורי שונות –**

* סעיף Craft (שאלות 21– 24) בעל שונות גבוהה, ייתכן כי הדובר נובע מהערכה שונה של סגנון כתיבה בין אדם למודל שפה.
* סעיף Stake Credibility ו Subplot Conflict קיבלו אצל המבקר האוטומטי ציונים מעט נמוכים, כנראה בגלל רגישות פחותה לפרטים לוגיים או דרמטיים.

המבקר האוטומטי ביצע הערכה עקבית מאוד לסיפור, ובמקרים בהם היה פער הוא נבע לעיתים מחוסר פרשנות רגשית מצד המבקר האוטומטי(לא "נסחף" ברגעי שיא כמו מבקר אנושי), או מהקפדה יתרה על אלמנטים טכניים (כמו אמינות לוגית או קצב עלילתי).

**סיפור 3 - Ashes of Vengeance**

**אזורי דמיון –**

* שאלות על זיהוי מטרות, התפתחות והקשר לדמויות משנה קיבלו ציונים דומים.
* שאלות על לוגיקה ועלילה עקבית – המבקר האוטומטי תאם לחלוטין למבקר האנושי.

**אזורי שונות –**

* פרשנות שונה לעומק השינוי הפנימי של הדמות הראשית – המבקר האנושי ראה שינוי מוגבל, בעוד שהמבקר האוטומטי פירש אותו כטרנספורמטיבי.
* תיאורים ספרותיים – שאלות Craft קיבלו ציונים גבוהים יותר מהמבקר האנושי.

המבקר האוטומטי עמד יפה במשימה, אך לעיתים מעניק קרדיט יתר על המידה כאשר מופיע שינוי סמלי או כללי מבלי שהוא מפורט לעומק.

**סיפור 4 - The Price of Redemption**

**אזורי דמיון –**

* זיהוי המבנה של הקונפליקט המרכזי והלוגיקה הכללית של רצף העלילה.
* שאלות על מבנה כמו כמות דמויות והתאמה לקו עלילה קצר – קיבלו ציון זהה.

**אזורי שונות –**

* פרשנות רגשית: המבקר האנושי התרשם יותר מתיאורים, בעוד המבקר האוטומטי מצא אותם מחוסרי פירוט.
* שאלות על סיכונים או רלוונטיות הקונפליקט קיבלו תוצאות שונות מעט – כנראה בגלל התייחסות טכנית מול חוויה אנושית.

המבקר האוטומטי העריך נכונה את המבנה והעלילה, אך פספס לעיתים ניואנסים ספרותיים שמבקר אנושי שם אליהם לב כמו הקשר בין תיאורים לדרמה.

**סיפור 5 - Ashes of Vengeance**

**אזורי דמיון –**

* רצף לוגי תקין והיעדר סתירות – ניתוח עקבי בין המבקר האנושי למבקר האוטומטי.
* קונפליקט מרכזי ואירועים ברורים.

**אזורי שונות –**

* העדר עומק רגשי – המבקר האנושי זיהה סתמיות במניעים בעוד המבקר האוטומטי פירש אותם כמספקים.
* שאלות על Craft – הבדל של עד 0.0375 בין ציון המבקר האנושי למבקר האוטומטי.

המבקר האוטומטי נותן יתר קרדיט על קיום מבנה פורמלי גם כשהתוכן הרגשי והעמוק חסר.

**סיפור 6 - Divided Hearts**

**אזורי דמיון –**

* הקונפליקט הראשי וההיגיון בין המבקרים היו עקביים.

**אזורי שונות –**

* פרשנות לתיאורים רגשיים הייתה זהירה יותר מצד המבקר האוטומטי.

המבקר האוטומטי נצמד למבנה הסיפור והקונפליקט, אך נוטה לפספס או לעדן את החוויה הרגשית של הדמויות. ההתאמה בכל זאת גבוהה.

**סיפור 7 – Empty Story**

* המבקר האוטומטי נתן אפס בכל שאלה בצורה עקבית כיוון שזיהה שקיבל טקסט ריק.
* המבקר האוטומטי הדגים כי עמד במקרה קיצון בו קיבל טקסט ללא תוכן, עם זאת הוא לא ניסה "לדמיין" או להעריך מתוך שורות ריקות וכתוצאה מכך נתן שיפוט מדויק לחלוטין.

**סיפור 8 - The Blue Hat**

* המבקר האוטומטי מזהה נכון את קיום מבנה בסיסי ונותן ניקוד לסדר תחבירי, לשפה תקינה ולהיעדר סתירות.
* כמעט כל יתר הקריטריונים נכשלים – אין קונפליקט, אין שינוי, אין עומק רגשי.
* המבקר האוטומטי מעניק נקודות היכן שמגיע, עבור שפה ודקדוק, אך מבקר בחומרה את חוסר העומק. הציון 1.53 משקף במדויק את הכוונה לכתוב סיפור רע. כלי הביקורת מגיב כראוי גם לקצוות.

**מסקנה מההשוואה:**

מהשוואת הדירוגים עולה כי הביקורת של המבקר האוטומטי מדגימה עקביות גבוהה והתאמה לביקורת האנושית, במיוחד בניתוח דמויות, קונפליקטים ולוגיקה נרטיבית. עם זאת, במרבית הסיפורים ניכרת שונות מסוימת בשאלות הנוגעות לעומק רגשי, סגנון כתיבה ופרשנות ספרותית, שם למבקר האנושי הייתה נטייה להעניק ציונים גבוהים יותר. במקרי קצה, כמו סיפור ריק או סיפור כתוב באיכות נמוכה בכוונה, המבקר האוטומטי זיהה את המגבלות באופן מדויק ושיקף את חולשתם בציון נמוך, מבלי להיסחף למחמאות או ניסיונות פרשנות מיותרים.

**בנוסף לדו"ח זה, מצ"ב תיקייה – Reports המכילה את כל הדירוגים שקיבלו הסיפורים על ידי המבקר האוטומטי ותיקייה נוספת – Stories המכילה את כל הסיפורים שנבחנו כקבצי טקסט.**