## **מדריך: כך תתמודדו עם "הזיות" של בינה מלאכותית בעבודה (ובשימוש אישי גם כמובן)**

## מאת: אייל מרקוס

## **המדריך הזה מספק טכניקות ונהלי עבודה לצמצום ההשפעה של הזיות בינה מלאכותית בהקשרים מקצועיים ואישיים כאחד.**

### **כלים ומתודולוגיה**

## השתמשתי בחמישה כלים של בינה מלאכותית וכלי בינה אנושית אחד:

## Perplexity - לתחקירים, לאיסוף והצגת הטקסט.

1. Claude - לכתיבה ולסידור הטקסט.
2. ChatGPT - לכתיבה של חלק מהטקסט.
3. Gemini - לתרגום.
4. Stealth GPT - להפיכת הטקסט לאנושי יותר.
5. **בינה אנושית - השתמשתי בכלי בינה אנושית אחד (שלי).**

הבינה האנושית שלי נתנה בריפים לכלים, ערכה, סיננה, שכתבה ולמעשה היא זו שיצרה את המדריך. הבינה המלאכותית עזרה לי מאוד, נחמד מצידה בסך הכל.

### **הקדמה**

### ״הזיות״, המדענים קוראים לזה.

## Hallucinations, Confabulations - אלה הכינויים הרווחים לעובדה שמודלי השפה דוגמת צ׳אטג׳פט, ג׳מיני, קופיילוט וקלוד "משקרים" בביטחון מוחלט.

## זה יכול להיות לינק מומצא, מחקר שלא קרה או אפילו מידע לא נכון מתוך מסמך שהעלתם אל הכלי.

## מה שנקרא - You had one job?!

## זה אתגר משמעותי, במיוחד בעולם המקצועי.

## אם אי אפשר לסמוך על התוצאות, זה מקשה מאוד על העבודה.

תסתכלו מסביב - ההזיות של הכלים הולידו שכבה שלמה של אנשים שלא מוכנים לגעת בכלים בגלל התוצאות הלא נכונות.

קודם כל - זו בעיה מובנית בכלים האלה. רבות מהסיבות קשורות ב״אימון״ של המודלים. אימון = המידע שהמודלים עיכלו וההנחיות המובנות במודלים (לאיתור מידע רלוונטי ועיבודו).

**מודלי שפה גדולים (LLMs) מייצרים ״הזיות״ מכמה סיבות:**

1. סטייה ממקורות מידע: נתוני האימון עשויים להכיל אי-דיוקים או סתירות, מה שמוביל ליצירת טקסט שסוטה מהמציאות.
2. מאגרי מידע חלקיים או סותרים: LLMs מסתמכים על מאגרי מידע עצומים שעשויים להכיל מידע חלקי או סותר, וזה משפיע על התשובות שלהם.
3. הנחיות עמומות או לא מספקות (שימו לב לנקודה הזו): כאשר המודל נתקל בפרומפט לא ברור, הוא עלול לפנות לניחושים המבוססים על דפוסים שלמד.
4. מגבלות נתוני האימון: LLMs מוגבלים לידע הקיים בנתוני האימון שלהם וחסרים גישה למידע חיצוני בזמן אמת.
5. אופי סטטיסטי של יצירת טקסט: LLMs מייצרים טקסט על בסיס דפוסים סטטיסטיים, מה שלעתים יכול להוביל לפלט שגוי עובדתית או חסר היגיון. אני לא אצלול לעומק במאמר הזה לאיך המודלים יוצרים טקסטים - רק זכרו שמדובר ביצירה על בסיס סטטיסטיקה. בהפשטה עצומה - הם Autocomplete משוכלל (מאווווד).
6. חוסר ביכולות הסקה: בעוד ש-LLMs יכולים לעבד וליצור טקסט, הם לא יכולים לעשות היסקים לוגיים או להבין היגיון בריא.
7. חוסר יכולת להבחין בין עובדה להמצאה: LLMs אינם יודעים באופן טבעי להבחין בין מידע עובדתי לבדיוני בנתוני האימון שלהם.
8. מורכבות השפה הטבעית: הניואנסים והעמימויות בשפה האנושית יכולים להוביל לפרשנויות שגויות ופלט לא נכון על ידי LLMs.

ועכשיו, לסיבה שבגינה התכנסנו כאן.

## **טכניקות לצמצום הזיות:**

פירטתי מגוון טכניקות - כולן בידי המשתמשים. יש טכניקות לצמצום הזיות שמגיעות ״עוד קודם״ בשלב בניית ואימון המודלים.

| **טכניקה** | **תיאור מפורט** | **דוגמה לפרומפט (מקצועי)** |
| --- | --- | --- |
| פרומפטים ברורים וספציפיים | זה כלל הבסיס לכל מלאכת פרומפטינג:יצירת פרומפטים עם הוראות מדויקות ומגבלות מוגדרות היטב מסייעת להנחות את התגובה של הבינה המלאכותית באופן מבוקר יותר. על ידי הגדרת התבנית, התוכן והמגבלות - תוכלו להפחית את נטיית הבינה המלאכותית "למלא את החסר" עם מידע פוטנציאלי לא מדויק. הטכניקה זו שימושית במיוחד כאשר אתם זקוקים למידע ספציפי, ממוקד ללא פרטים מיותרים. | "צור תיאור מוצר עבור הסמארטפון החדש XYZ-5000 שלנו. השתמש רק במפרט הטכני המסופק במסמך המצורף. התיאור צריך להיות באורך של 150-200 מילים, אנא הדגש את 3 התכונות העיקריות ותמנע מהשוואה למוצרים של המתחרים." |
| Retrieval-Augmented Generation (RAG) | RAG היא טכניקה עוצמתית המשלבת את היכולות היצירתיות של מודלי השפה (צ׳אטג׳פט ודומיו) עם היכולת לאחזר ולהטמיע מידע חיצוני. שיטה זו מאפשרת לבינה המלאכותית לגשת ולהשתמש במידע מאומת ומעודכן ממאגר ידע שאתם תספקו לה, ובכך מפחיתה באופן משמעותי את הסבירות להזיות (תחשבו למשל על בניית GPT משלכם, עם מסמכים ארגוניים שתעלו אל ה-GPT). על ידי קישור התשובות לנתונים עובדתיים שניתן לאחזר, RAG מבטיח שהפלטים של הבינה המלאכותית יהיו מדויקים ואמינים יותר. | "באמצעות המידע הכלול אך ורק במדריך הרשמי לעובדים של החברה שלנו (גרסה 2024.1), אנא כתוב אימייל המסביר לצוות העובדים את מדיניות החופשה שלנו להורים. כלול פרטים על זכאות, משך זמן והליך הבקשה לחופשה." |
| Chain of Thought Prompting - שרשרת המחשבה | טכניקה מתקדמת זו של כתיבת פרומפטינג מעודדת את הבינה המלאכותית לפצל בעיות מורכבות לצעדים לוגיים קטנים יותר (אגב, זה מעולה גם עם בני אדם, לא במקרה). על ידי הדרכת המודל דרך תהליך חשיבה - שלב אחר שלב, יש פחות סיכויים שהמודל יסיק לוגית נתונים שיכולים להוביל להזיות. השיטה הזו יעילה במיוחד עבור משימות הדורשות חשיבה במספר שלבים או חישובים, שכן היא מאפשרת מעקב ואימות טובים יותר של תהליך המחשבה של הבינה המלאכותית. בנימה אישית - זה תהליך מרתק של פרומפטינג, תנסו בבקשה. | "חשב את ההכנסות הצפויות של החברה שלנו לרבעון השלישי של 2024. הצג את ההיגיון שלך שלב אחר שלב, כולל: 1) ניתוח הביצועים ההיסטוריים ברבעון השלישי, 2) התחשבות במגמות השוק הנוכחיות, 3) השפעת השקות המוצרים החדשים שלנו, ו-4) כל גורמים עונתיים. הסבר כל שלב בחישוב שלך." |
| הפרומפט השלילי - Negative Prompting | טכניקה זו כוללת הצהרה מפורשת של מה שלא צריך להיות כלול בתגובה של הבינה המלאכותית. על ידי זיהוי והדרה פרואקטיבית של מקורות פוטנציאליים של הזיות או תוכן לא הולם, תוכלו להנחות את הבינה המלאכותית לכיוון פלטים מדויקים ורלוונטיים יותר. זה שימושי במיוחד כאשר מתמודדים עם נושאים רגישים או כאשר ישנם סוגים ספציפיים של מידע שיש להימנע מהם. | "כתוב בבקשה טיוטת הודעה לעיתונות המסבירה את נתוני הרבעון הרביעי של החברה שלנו. אל תכלול תחזיות לעתיד, תחזיות של הכנסות או אזכורים של מוצרים שלא פורסמו. אנא היצמד למידע עובדתי בלבד על הביצועים ברבעון האחרון, לפי הדו״ח הכספי שאספק לך." |
| תבניות נתונים מובנות | מתן מבנה ברור או תבנית לפלט של הבינה המלאכותית יכול להפחית באופן משמעותי את הסבירות להזיות על ידי הגבלת פורמט התגובה. טכניקה זו יעילה במיוחד כאשר מתמודדים עם הצגת נתונים או כאשר פורמט עקבי הוא קריטי. על ידי ציון המבנה המדויק הצפוי, תצליחו להגביל את יכולתו של הכלי להציג מידע מיותר או פוטנציאלי לא מדויק. | "הצג את נתוני המכירות שלנו עבור הרבעונים הראשון-רביעי של 2023 בטבלה עם העמודות הבאות: רבעון, הכנסה (במיליוני דולר), צמיחה רבעונית (גדילה באחוזים), והמוצר הנמכר ביותר. כלול שורה עבור כל רבעון ושורה של ״סה"כ״ בסוף כל עמודה." |
| פעולה לוגית בודדת | על ידי הגבלת הבינה המלאכותית לבצע פעולה לוגית אחת או לענות על שאלה ספציפית אחת בכל פעם, אתה מפחית את מורכבות המשימה ובכך את הפוטנציאל לשגיאות או להזיות. גישה זו שימושית במיוחד להפקת מידע מדויק או כאשר מתמודדים עם נושאים מורכבים שניתן לפצל לשאילתות פשוטות יותר, בודדות. | "על סמך הדוח המכירות שאעלה, מה היה המספר המדויק של היחידות שנמכרו עבור המוצר הדגל שלנו, XYZ-5000, בצפון אמריקה במהלך הרבעון השני של 2023?" |
| סקירה אנושית ובדיקת עובדות | זו אמנם לא טכניקה למניעת הזיות של המודל, אבל חובה להכניס לכל תהליך עבודה עם בינה מלאכותית סקירה אנושית - כדי לעלות על אי-דיוקים (וכמובן לתקן אותם). סקירה אנושית היא קריטית במיוחד לתקשורת רגישה או בעלת סיכונים גבוהים. | "צור טיוטה ראשונה של המכתב השנתי למחזיקי המניות שלנו. התמקד בהישגנו המרכזיים, במדדי הפיננסיים ובחזון האסטרטגי שלנו. תייג כל קטע שבו אתה לא בטוח לגבי מספרים או טענות ספציפיים, כך שהצוות המנהל שלנו יוכל לאמת ולמלא את החסר במהלך הסקירה." |
| בקשו מהמודל - לענות רק אם יש לו את התשובה המדויקת | אנת׳רופיק, החברה שמאחורי קלוד, [ממליצה לתת לכלי סולם לרדת ממנו](https://github.com/anthropics/courses/blob/master/prompt_engineering_interactive_tutorial/Anthropic%201P/08_Avoiding_Hallucinations.ipynb). כלומר? להנחות את הבינה המלאכותית לספק רק מידע שהיא בטוחה בו, ולהביע במפורש אי ודאות כאשר היא לא בטוחה במידע. הטכניקה הזו מעודדת את המודל להיות זהיר יותר בתגובותיו ולהבחין בבירור בין מידע עובדתי לבין ספקולציות פוטנציאליות.  | "ענה על השאלה הבאה לגבי לוגיסטיקת שרשרת האספקה שלנו: 'מהו זמן המשלוח הממוצע שלנו מהמפעל שאנגחאי למרכז ההפצה שלנו בלוס אנג'לס?' אם אין לך גישה למידע מדויק ומעודכן על כך, אנא השב ב: 'אין לי נתונים מדויקים כדי לענות על שאלה זו. מידע זה צריך להיות מאומת עם מחלקת הלוגיסטיקה שלנו.'" |
| איסוף עובדות | גישה זו בשני שלבים כוללת תחילה הנחיה לבינה המלאכותית להפיק מידע רלוונטי ממקור נתון, ולאחר מכן שימוש רק במידע המופק זה כדי לנסח את תגובתה. על ידי הפרדה מפורשת של שלבי איסוף המידע ויצירת הפלט, ניתן להבטיח שהפלט של הבינה המלאכותית יהיה מבוסס ישירות על חומר המקור, ובכך להפחית את הסיכון להצגת מידע חיצוני או הזוי (כלומר - קודם בקשו מהכלי להציג את המידע מהמסמך, ורק אז בקשו לעבוד על המידע שהוא שלף). | "משימה 1: הפק את כל ההצהרות הקשורות ליוזמות השונות והכוללות של החברה שלנו מדוח האחריות החברתית של 2023. משימה 2 (פרומפט המשך): בהתבסס אך ורק על ההצהרות שאספת מהדו״ח, סכם את המטרות העיקריות שלנו בתחום השונות בחברה וכלול את כל הפעולות הספציפיות שנקטנו כדי להשיג אותן בשנת 2023. אל תכלול כל מידע שאינו נתמך ישירות על ידי המידע שאספת מהמסמך." |

### **מקורות:**

### [OpenAI](https://platform.openai.com/docs/guides/prompt-engineering) - מדריך פרומפטינג

### [אנת׳רופיק](https://github.com/anthropics/courses/blob/master/prompt_engineering_interactive_tutorial/Anthropic%201P/08_Avoiding_Hallucinations.ipynb) - מדריך פרומפטינג

### [זאפייר](https://zapier.com/blog/ai-hallucinations/)

### [turing.com](http://turing.com)

### **כמה מילות סיכום:**

### התוכן שמופק על ידי בינה מלאכותית צריך לשמש כנקודת התחלה או אמצע - הוא לא התוצר הסופי. המוח שלכם.ן הוא זה שצריך ליצור תוצר סופי.

### חשוב להבין את המגבלות (זה ישתנה ובשנים הקרובות - ההזיות יצומצמו מאוד).

### צפו לפגמים: אל תצפו לתוצאות מושלמות מה-AI. צפו לטעויות, הזיות ולפעמים אפילו תשובות מוזרות.

### חשוב מאוד שהבינו מדוע התוצאות מוזרות: גם פרומפטים מנוסחים היטב עלולים לגרום לכלי ה-AI לייצר הזיות. גם הטכניקות שפירטתי למעלה לא תמיד יעבדו. המודלים הם סטטיסטיים, וקשה לצפות את אופן התשובה ואת חוסר העקביות.

התמודדות עם הזיות בינה מלאכותית דורשת גם הבנה של כיצד עובדים המודלים, וכיצד להוציא מהם את התוצאות הטובות ביותר.

אם נבין את האופן שבו מערכות בינה מלאכותית פועלות, ניישם טכניקות תקשורת ברורות, נאמת בקפדנות את המידע, נוכל לצמצם משמעותית את הסיכון להזיות ולשפר את האמינות והערך של התוצאות שתקבלו מהעבודה עם כלי ה-AI.

**בהצלחה גדולה!**