המאמר היומי של מייק - 09.12.24:  
Scaling Synthetic Data Creation with 1,000,000,000 Personas

תמצית המאמר ותרומות מרכזיות:

1. השקת Persona Hub:

- מאגר של מיליארד פרסונות מגוונות שנוצרו באמצעות טכניקות הניתנות להרחבה

- פרסונות אלו מגלמות ידע, תחומי עניין, התנסויות ומקצועות ייחודיים, המייצגים כ-13% מאוכלוסיית העולם

2. יצירת דאטה סינתטי מבוסס פרסונות:

- שילוב פרסונות בפרומפטים מאפשר למודלי שפה גדולים (LLMs) לייצר נתונים סינתטיים מגוונים במיוחד

- מדגים יישומים במגוון תחומים כגון בעיות מתמטיות, חשיבה לוגית, הוראות, טקסטים עתירי ידע, דמויות NPC במשחקים וממשקי כלים

3. שיטות ליצירת פרסונות:

טקסט-לפרסונה:

- מייצר פרסונות ישירות מנתוני רשת- מנתח הקשר טקסטואלי כדי להסיק את הפרסונה שסביר שקשורה אליו (למשל, "מי עשוי לכתוב או לחבב טקסט זה?")

- מפיק תיאורי פרסונה גסים או מדויקים (למשל, "מדען מחשב" לעומת "חוקר למידת מכונה המתמקד בארכיטקטורות נוירונים")

- מתרחב בקלות באמצעות LLMs ומאגרי נתונים ציבוריים ענקיים

פרסונה-לפרסונה:

- מרחיב פרסונות באמצעות קשרים יחסיים (למשל, ילד הקשור לאחות ילדים, או קבצן הקשור לעובד מקלט)

- משתמש בפרומפטים מבוססי יחסים כמו "מי נמצא בקשר קרוב עם פרסונה זו?"

- העשרת פרסונות נוספת על ידי איטרציה של שש דרגות הפרדה

4. תהליך הסרת כפילויות פרסונות:

- MinHash Deduplication: מסיר פרסונות דומות על בסיס חפיפת n-gram טקסטואלית

- Deduplication מבוסס אמבדינג: מסנן פרסונות באמצעות דמיון סמנטי (מרחק קוסיין) המחושב דרך אמבדינגים. ספי הדמיון הותאמו בהתאם לשיקולי איכות מול כמות

- לאחר ניקוי והסרת כפילויות, המאגר כלל 1,015,863,523 פרסונות ייחודיות

5. יישומים:

א. סינתזת בעיות מתמטיות:

- יצר 1.09 מיליון בעיות מתמטיות ייחודיות באמצעות פרסונות

- מודל 7B שעבר טיוב (fine-tuning) עדין עם בעיות אלו השיג דיוק של 79.4% על סט בדיקה סינתטי תוך-התפלגות ו-64.9% על MATH, תוצאה המשתווה ל-GPT-4-turbo-preview  
 - הדגים יכולת הרחבה - הוספת פרסונות שיפרה את גיוון הבעיות והבטיחה כיסוי רחב של מושגים מתמטיים

ב. בעיות חשיבה לוגית:

- סינתז חידות לוגיות מאתגרות (למשל, חשיבה מרחבית או זמנית) המותאמות למאפייני פרסונה

- כלל בעיות בסגנון Ruozhiba שובבי לבדיקת יכולות לוגיות מעודנות

ג. יצירת הוראות:

- יצר שאילתות משתמש המשקפות פרסונות מגוונות מהעולם האמיתי (למשל, כימאי עשוי לבקש מערכי ניסוי; אמן עשוי לבקש טכניקות ציור)

- אפשר סימולציות של שיחות רב-שלביות בין משתמש ל-LLM על ידי שרשור פרומפטים של פרסונות

ד. טקסטים עתירי ידע:

- יצר מאמרים ותוכן חינוכי המתואמים עם מומחיות הפרסונות (למשל, גנן כתב מדריכים על צמחים עמידים לבצורת)

- כיסה כמעט כל נושא באמצעות הרוחב של הפרסונות

ה. פיתוח כלים (פונקציות):

- חזה כלים שפרסונות עשויות להזדקק להם (למשל, נהג מונית הזקוק ל-API של תנאי תנועה)

- יצר הגדרות כלים עם קלטים, פלטים ותלויות ברורים

6. תוצאות מרכזיות:

- מודלים קטנים יותר (למשל, Qwen2 7B) שעברו כוונון עדין באמצעות נתונים סינתטיים השיגו רמות ביצועים שבדרך כלל דורשות מודלים גדולים יותר

- הוכיח שגיוון פרסונות מוביל לפלטים מגוונים ויצירתיים משמעותית יותר  
- הדגים שפרסונות יכולות לדמות התנהגויות משתמש מגוונות, ולפעול ביעילות כנושאות מבוזרות של זיכרון ה-LLM

7. סיכום  
המאמר מסמן קפיצת מדרגה (לא ברור עד כמה משמעותית) בגנרוט דאטה סינתטי. המתודולוגיה המוצעות נראית מבטיחה וניתנת ליישום עבור מגוון משימות, ויוצרת הזדמנויות לטיוב חכם של LLM, פיתוח יישומים, ואפילו סימולציות חברתיות.

https://arxiv.org/abs/2406.20094