

בנוסח מחשבתי:

למודלי שפה קיימת מגבלת context window ולכן לא ניתן להזין אליהם בבת אחת כמויות גדולות של מידע גולמי כמו 10,000 חשבוניות. תשובה לשאלה "כמה שילמנו לספק X בשנת 2023?" דורשת עיבוד של מסמכים רבים ולא ניתן להכניס את כולם לLLM וזה גם לא נכון ארכיטקטוני. הפתרון הוא להפריד בין אחסון וחישוב נתונים ולבין הבנה והצגה בצורה טבעית. כל החשבוניות נשמרות כנתונים במסד הנתונים עם השדות: vendor, invoice date, amount, currency, line items. אחסון זה יכול לתת לנו את האפשרות לבצע שאילתות SQL על מסד הנתונים. למשל נוכל לעשות aggregation לפי ספק ושנה ונחשב את הסכום הכולל. LLM לא יקבל חשבונית אלא את תוצאת החישוב + שאלה מקורית של המשתמש ותפקידו יהיה לנסח תשובה בשפה טבעית ולהציג את המידע בצורה ידידותית למשתמש.