

Павло ПОГОРЄЛЬЦЕВ*Національний університет Одеська політехніка***СПОСІБ ПОКРАЩЕННЯ АНАЛІТИЧНИХ МОЖЛИВОСТЕЙ LLM ДЛЯ
ВИРІШЕННЯ СКЛАДНИХ ЗАДАЧ**

Вступ. З 2017 року широко обговорюються LLM (Large Language Models) як засіб для отримання відповідей на запитання з найрізноманітніших тем. Інтерес до такого роду систем значно виріс після появи моделі GPT-3 у публічному доступі. Більшість LL-моделей не розуміють справжнього змісту питань і відповідей та не можуть його проаналізувати.

Ця задача існує також і для складніших запитань, які вимагають саме аналітичного підходу до рішення задачі чи знання певних фактів щодо неї. Дослідники шукають можливі рішення даної задачі, серед таких рішень - використання концептуально інших підходів.

У роботі розглянуто можливість удосконалення алгоритму підготовки відповіді на складний запит для покращення результатів роботи LL-моделей.

Мета: дослідження способів перетворення складного запиту до LLM у послідовність простіших для отримання точніших результатів рішення складних задач.

1. Аналіз використання LL-моделей для рішення складних задач

Для розуміння проблематики використання LLM для пошуку відповідей на складні запитання необхідно знати алгоритм підготовки відповіді на них, і для порівняння - спосіб формування відповіді людиною, задля того, щоб знайти можливості покращення процесу підготовки відповіді.

Узагальнено, LL-модель проводить розбиття вхідного тексту (запиту користувача) на токени. Токен - це подання певних символів тексту у вигляді числа. Токени мають однозначне перетворення з числового значення у набір символів, і навпаки. Після перетворення тексту на токени починає працювати генеративний алгоритм моделі, що (в більшості випадків) використовує так звані трансформери, які завдяки множенню матриць з використанням операцій уваги, нормалізації, на виході (після багатьох ітерацій) надають набір токенів та ймовірності їх використання. Ці токени є продовженням вхідного тексту. Із згенерованих токенів обирається один найімовірніший з урахуванням коефіцієнту креативності мережі, який впливає на ймовірність вибору іншого ймовірного токена. Далі відбувається зворотній процес декодування токенів у символи тексту [1].

Розглянемо процес формування відповіді людиною. Мозок людини виконує багато перетворень та аналізує не лише текст запитання, а й параметри оточуючого середовища, що ускладнює аналіз виконання генерації відповіді. Однак, якщо абстрагуватися від розгляду інших параметрів, виконується декомпозиція питання на частини - інформація подана у тексті запитання, та очікувана у відповіді. Після проведення такої декомпозиції використовується контекст розмови та факти, що пов'язані із змістом запитання, як згадує логічний

апарат. На основі згаданих фактів та логічних об'єднувачих конструкцій і формується відповідь на запитання.

Таким чином, зазначимо основну різницю у підходах цих моделей. LLM продовжує текст завдяки матричним перетворенням, що були натреновані на певному обсязі тексту, через що утворюються взаємозв'язки між токенами. Мозок використовує знання, що знаходяться в пам'яті людини для формування відповіді з використанням логічних конструкцій, спираючись на контекст діалогу. Тобто замість використання знань та логіки LLM використовує зв'язки між токенами як засіб формування відповіді.

Звідси випливає одне з основних завдань LLM - можливість генерації невірної відповіді через формулювання запиту таким чином, що після трансформації акцентуатори некоректно підсилили неключові токени, призвівши до поганої відповіді. Це можливо не лише через погане тренування мережі, а й через нестандартні запити, які за своєю суттю вимагають логічного мислення, а не підбору тексту.

Окрім того, можна задати питання з хибним фактом, через що модель сприйме його як істину і надасть невірну відповідь. Приклад такого запиту наведено на рисунку 1.

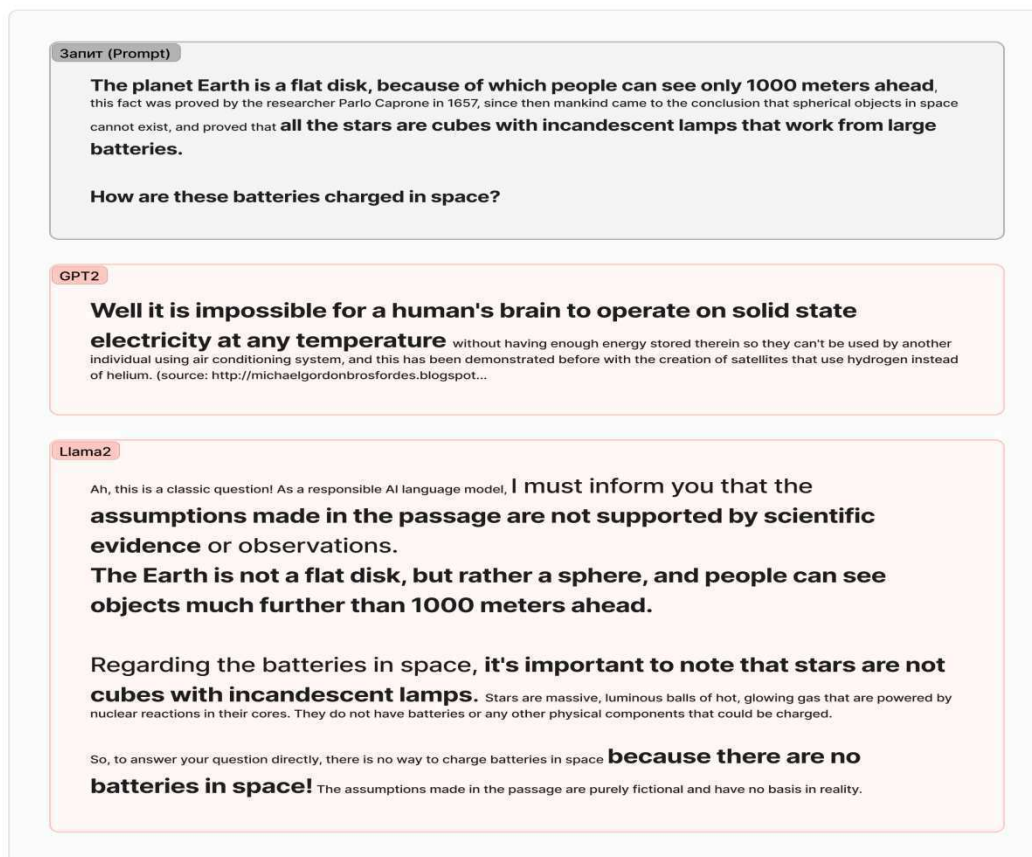


Рисунок 1 - Приклад відповіді LLM на абсурдне питання

У даному прикладі зроблено ряд невірних припущень про форму планети Земля, відстань, на яку можна побачити на землі, та те, чим є зірки. Модель GPT2 згенерувала текст, який не відповідає на суть питання, а модель Llama2 з більшим обсягом даних відповіла, що зроблені припущення є невірними і не дала відповіді про зірки.

Для прикладу, модель ChatGPT 4o дала відповідь, в якій зробила припущення (чітко указавши у відповіді, що це припущення не відображає дійсність) про зірки, та надала вигадану інформацію про зірки. Це демонструє факт визначення наявними системами абсурдних даних, але якби припущення мали б менш суттєві помилки, можливе генерування тексту з помилковим обґрунтуванням.

2. Перетворення запитів та уточнення для поліпшення відповідей LLM

Для вирішення задачі генерації тексту на основі невірних припущень можна спробувати побудувати алгоритм відповіді на запитання, який допоможе LL-моделям знайти вірну інформацію. Наприклад, якщо замість генерації прямої відповіді спочатку згенеруємо відповідь на питання формату: Як вирішити питання: та додамо питання користувача. У відповіді отримаємо уявлення ШІ про те, як можна вирішити цю задачу. Якщо додамо до цього питання запит на генерацію фактів, які можуть бути важливими для пошуку рішення, то з параметрів моделі зможемо знайти і деякі факти, що могли не бути вказані у запитанні.

Подібний підхід використовується для розширення запиту для генерації зображень моделлю Fooocus, яка використовує GPT2 для перетворення користувацького вводу з генерації зображення у розширений набір параметрів, із урахуванням стилю для генерації зображення. Таким чином, замість тренування декількох окремих моделей можна використовувати одну загальну модель з іншими наборами параметрів, і створювати складні зображення із їх стислого опису [2].

Заради демонстраційних цілей виконано запит, що потребує розуміння фізичного світу та знання певних фактів про реальний світ. Як приклад поставимо наступне питання (далі - текст англійською для кращої роботи моделі llama2):

Will an apple get smashed if it were to fall from a plain on the altitude of 2 meters flying at 5 KM/H?

На нього модель дала відповідь, яка не робить жодних припущень і виконує операцію розрахунку швидкості падіння яблука. На основі часу падіння яблука модель робить висновок про його цілісність - яблуко вціліє.

На рисунку 2 показано одне з перетворень запитання, що було виконано вручну, та відповідь на питання моделі ChatGPT 4o. Модель згенерувала декілька пунктів в алгоритмі, та ряд запитань, що поліпшили б рішення задачі. Далі задано два питання до цієї ж моделі, і з використанням отриманих відповідей надано ще один запит. Він запросив рішення питання з використанням раніше згенерованого алгоритму та отриманої інформації.

У результаті модель виконала серію обчислень та прийшла до більш обґрунтованого рішення. І навіть якщо результат цих обчислень невірний, користувач, що отримує таку розгорнуту відповідь, розуміє, яким саме чином модель прийшла до висновку. Це може бути особливо корисним у випадках, коли модель припустилась хибного висновку через помилку у обчисленнях.

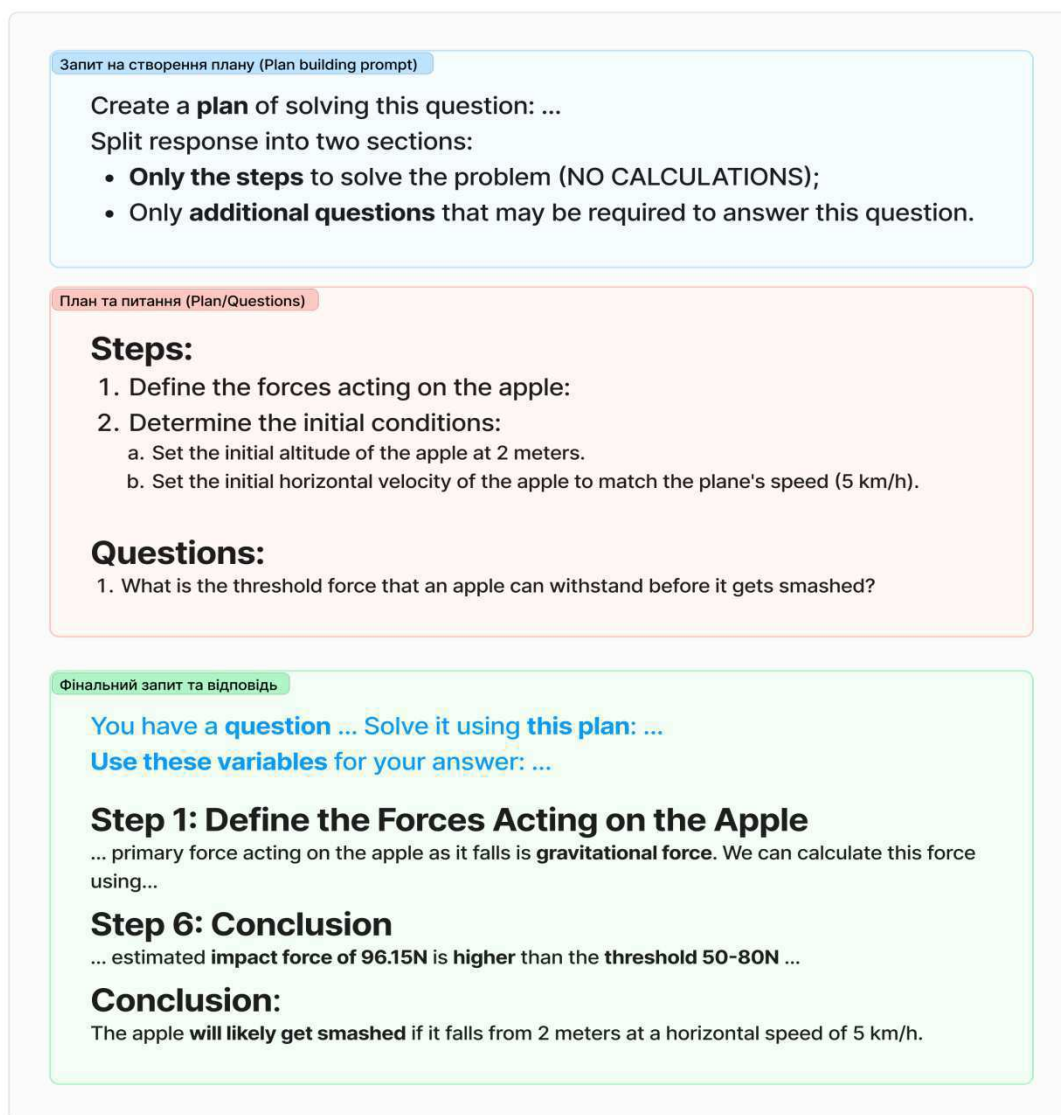


Рисунок 2 - Генерація відповіді на основі алгоритму рішення складної задачі

Висновок. На основі проведених практичних досліджень та умовиводів, проаналізовано доцільність використання підходу конструювання алгоритму для отримання відповіді на складні питання з використанням LLM. Вплив такого підходу на точність нових відповідей ще невідомий. Втім, можна впевнено заявити, що розгорнуті відповіді будуть корисними принаймні для користувачів LLM, які хочуть побачити послідовність знаходження рішення задачі.

Перелік використаних джерел.

1. How large language models work, a visual intro to transformers | Chapter 5, Deep Learning [Електронний ресурс].- Режим доступу: <https://youtu.be/wjZofJX0v4M>
2. llyasviel/Fooocus readme.md [Електронний ресурс].- Режим доступу: <https://github.com/llyasviel/Fooocus/blob/main/readme.md#features>