Харківський національний університет імені В.Н. Каразіна
Факультет математики і інформатики
Кафедра прикладної математики

Кваліфікаційна робота

магістра

на тему «*Інтеграція моделей штучного інтелекту в SAS для аналізу даних клінічних випробувань*»

Виконав:
студент групи МП-61
2 курсу
спеціальність 113 – прикладна математика
освітньо-наукова
програма «Прикладна математика»

Скляр І.К.

Науковий керівник: *проф., доктор техн. наук, Ромашов Ю.В.*

Науковий керівник з виробництва: *Principal Statistical Programmer and Analyst, магістр, Артемчук А.В.*

Рецензент: *Олексієнко Анна Сергіївна, Statistical Programmer/Analyst, "Intego Group, LLC", магістр за напрямом прикладної математики*

Харків - 2024 рік

АНОТАЦІЯ

Оскільки галузь клінічних випробувань продовжує розвиватися, зростає інтерес до використання технологій штучного інтелекту (ШІ) та обробки природної мови (NLP) для спрощення програмувальних завдань та підвищення ефективності. Однією з перспективних технологій у цій сфері є ChatGPT, розширена модель мови, яка показала вражаючі результати у різноманітних застосуваннях.

У кваліфікаційній роботі я дослідив потенціал інтеграції ChatGPT у робочі процеси програмування SAS для клінічних випробувань. Починаючи з архітектури і принців роботи ChatGPT, я зосередив увагу на інноваційному методі трансформерів моделі ШІ.

Далі я запропонував метод інтеграції моделі ШІ в середовище програмування SAS, розглядаючи конкретні задачі, з якими може допомогти ChatGPT, включаючи створення наборів даних ADaM на основі стандартів CDISC, генерацію фігур та таблиць за допомогою мови програмування SAS 9.4, генерацію тестових даних для клінічних випробувань.

Нарешті, була оцінена точність та ефективність ChatGPT у аналізі даних, спираючись на приклади з реального світу клінічних досліджень, спираючись на питання, пов'язаних з конфіденційною інформацією та іншими обмеженнями.

ABSTRACT

As the field of clinical trials continues to evolve, there is increasing interest in leveraging artificial intelligence (AI) and natural language processing (NLP) technologies to streamline programming tasks and enhance efficiency. One of the promising technologies in this area is ChatGPT, an advanced language model that has demonstrated impressive results in various applications.

In this paper, I explored the potential of integrating ChatGPT into SAS programming workflows for clinical trials. Starting with the architecture and working principles of ChatGPT, I focused on the innovative method of AI model transformers.

I then proposed a method for integrating the AI model into the SAS programming environment, considering specific tasks that ChatGPT can assist with, including creating ADaM datasets based on CDISC standards, generating figures and tables using the SAS 9.4 programming language, and generating test data for clinical trials.

Finally, I evaluated the accuracy and effectiveness of ChatGPT in data analysis, drawing on real-world examples from clinical research, addressing issues related to confidential information and other limitations.

ЗМІСТ

АНОТАЦІЯ.....2
ВСТУП.....4
1. Архітектура та принципи роботи NLP та GPT нейронних мереж.....5
1.1 Принципи роботи GPT нейронної мережі.....5
1.2 Архітектура трансформерів.....6
1.2.1 Pre-Training та Fine-Tuning.....7
1.2.2 Encoder та Decoder.....8
1.2.3 Механізм Уваги.....9
1.2.4 Генерація відповідей.....10
2. Інтеграція ШІ ChatGPT у середовище програмування SAS.....13
2.1 З'єднання ChatGPT через SAS 9.4 IDE.....15
3. Вирішення завдань за допомогою ШІ.....18
3.1 Перевірка стандартів за допомогою ChatGPT.....18
3.2 Обробка програмного коду штучним інтелектом для створення документації.....19
4. Генерація тестових даних за допомогою моделі ШІ ChatGPT.....22
4.1 Методи налаштування даних під час генерації.....24
4.2 Застосування моделі генерації даних у клінічних випробуваннях.....25
5. Врахування регуляторних обмежень та конфіденційності.....31
5.2 Обмеження регуляторних органів.....32
ВИСНОВКИ.....35
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....37

ВСТУП

У галузі досліджень клінічних випробувань управління даними є основою успішних результатів випробувань. Незважаючи на технологічний прогрес, сфера клінічних досліджень продовжує стикатися з викликами, які пов'язані з ручними та часозатратними завданнями, пов’язаними зі статистичним програмуванням та аналізом даних. Проблема полягає в рутинній логістиці управління даними, що в кінцевому підсумку призводить до збільшення витрат і затримки клінічних випробувань.

Більш того, на основі дослідження 2021 року Центру Тафтса з вивчення розробки лікарських засобів було оцінено [1], що середня вартість розробки рецептурного препарату, який успішно отримує схвалення на ринку, становить 2,6 мільярда доларів, де клінічні випробування є значною частиною цих витрат. Будь-яка невідповідність у процесі аналізу даних може призвести до затримок, завищення витрат та, у найгірших випадках, до порушення цілісності даних клінічних випробувань, що врешті-решт перешкоджає розробці нових лікувань.

Значний крок вперед в управлінні клінічними даними та методології програмування пов’язаний з випуском моделі штучного інтелекту ChatGPT®. ChatGPT використовує передові техніки машинного навчання (ML) та обробки природної мови (NLP), які роблять цю модель на базі ШІ здатною обробляти та генерувати текст, схожий на людський, і виконувати складні команди для управління даними та аналітиків даниї з небаченою швидкістю та гнучкістю.

Метою цієї роботи є дослідження інтеграції ChatGPT у робочий процес аналітика даних, дослідження задач, які може виконати ШІ, та генерації зразкових даних для клінічних випробувань. Автоматизуючи процеси, ChatGPT має потенціал значно зменшити час і зусилля, необхідні для рутинних процесів для програмування з використанням реальних даних, потенційно сприяючи швидшому процесу клінічних випробувань.

1. Архітектура та принципи роботи NLP та GPT нейронних мереж

1.1 Принципи роботи GPT нейронної мережі

Обробка природної мови (NLP) - це галузь на перетині інформатики, штучного інтелекту (ШІ) та лінгвістики. Її мета - дати можливість комп'ютерам розуміти, інтерпретувати та виробляти людські мови. Поява нейронних мереж значно просунула галузь обробки природньої мови NLP, дозволивши досягти прориву в таких завданнях, як класифікація текстів, машинний переклад і системи відповідей на запитання.

Моделі Generative Pre-trained Transformer (GPT), серія моделей NLP, розроблених компанією OpenAI, представляють собою значний крок вперед у цій галузі. Ці моделі базуються на архітектурі трансформера, яка стоїть в центрі багатьох недавніх успіхів у глибокому навчанні для NLP.

ChatGPT – це передова модель обробки мови, яка теж розроблена компанією OpenAI. Її призначення - генерувати текст, схожий на людський, що означає здатність розуміти та виробляти мову на подібний до нас манер. Це відбувається завдяки архітектурі нейронної мережі, яка називається «трансформер». Вона дозволяє моделі розуміти зв'язки між словами, фразами і навіть цілими реченнями в заданому контексті.

Тож чому аналітики даних у сфері клінічних випробувань повинні зацікавитися мовною моделлю? По-перше, ChatGPT може спростити програмні завдання. Вона може пропонувати фрагменти коду, надавати загальні рекомендації, а також допомагати у відлагодженні (debugging) коду. По суті, це високоінтелектуальний генератор тексту, натренований на величезному наборі даних, включаючи зміст, пов'язаний з програмуванням або аналізом даних. І хоча в нього немає вбудованих знань будь якої мови програмування, він достатньо розумний, щоб надавати допомогу в цих областях.

1.2 Архітектура трансформерів

ChatGPT є штучним інтелектом, який базується на архітектурі трансформерів, вперше представленій у 2017 році у науковій статті "Attention is All You Need" [2]. Суть архітектури трансформера полягає в тому, що вона оптимізує процес обробки послідовностей даних (наприклад, тексту), використовуючи механізм уваги, який дозволяє моделі зосереджуватися на релевантних частинах вхідних даних при генерації відповіді.

ChatGPT тренується на величезній кількості текстових даних за допомогою методу навчання під наглядом (кероване навчання), а також технік файн-тюнінгу з використанням зворотного зв'язку від людей. Це дозволяє моделі не тільки точно відповідати на запитання або продовжувати текстові фрагменти, але й адаптуватися до різних стилів спілкування, контекстів та вимог до змісту відповідей.

Принципи роботи ChatGPT можна розглянути на кількох рівнях:

- Попереднє навчання (Pre-training): На цьому етапі модель тренується на широкому спектрі текстових даних з метою зрозуміти мову, її структуру, зв'язки між словами та фразами. Використовуючи велику кількість текстів з різноманітних джерел, модель набуває здатності до генерування змістовного тексту.

- Файн-тюнінг (Fine-tuning): Після попереднього навчання модель може бути додатково навчена на специфічних даних або для виконання конкретних завдань. Цей процес дозволяє покращити якість відповідей моделі у певних доменах знань або стилях комунікації.

- Механізм уваги (Attention Mechanism): Ядром архітектури трансформера є механізм уваги, який дозволяє моделі "зосереджуватися" на важливих словах у вхідному тексті для кращого розуміння контексту та змісту запитань або коментарів.

- Генерація відповідей: На основі аналізу вхідних даних та з врахуванням навчених залежностей та контексту, модель генерує відповідь. Цей процес включає вибір наступного слова на основі ймовірностей.

Важливо зазначити, що ChatGPT був додатково налаштований для конкретних завдань або сфер застосування. У нашому випадку це означає, що його можна адаптувати для допомоги у аналізі для клінічних випробувань. Він навчався не лише на загальних текстах, але і на більш специфічному наборі даних, пов'язаних з клінічними дослідженнями. Це робить його більш здатним розуміти унікальну мову та вимоги цієї галузі.

Але треба пам'ятати, що ChatGPT - це інструмент, а не заміна людської експертизи. Хоча він може допомагати вам у різноманітних завданнях, важливо перевіряти його пропозиції. Особливо в клінічних випробуваннях, де дотримання стандартів та точність мають вирішальне значення.

Підсумовуючи, ChatGPT - це технологія, що постійно розвивається і удосконалюється. Його різні версії – починаючи від GPT-1 і до нинішнього ChatGPT – показали помітне поліпшення у можливостях генерації мови. Для аналітиків даних ця неперервна інновація пропонує перспективний шлях до підвищення ефективності та продуктивності у світі клінічних випробувань.

1.2.1 Pre-Training та Fine-Tuning

Процес попереднього навчання (Pre-training) у контексті моделей на кшталт ChatGPT та інших трансформерів не описується однією конкретною формулою, а скоріше є набором методів для навчання моделі на великому наборі неструктурованих даних перед тим, як виконувати фін-тюнінг на специфічному наборі даних або задачах. Основна мета попереднього навчання - розвинути глибоке розуміння мови, яке може бути адаптовано до різноманітних завдань обробки природної мови. [1]

Хоча конкретна формула попереднього навчання може змінюватись в залежності від використаного методу, основні методи можна представити через задачу прогнозування наступного слова. Наприклад, в моделі може використовуватись втрата перехресної ентропії для оцінки її здатності правильно передбачати наступне слово:

$$L = -\sum_{i}^{N} \ln p(\omega_i | \omega_{i-1}, \omega_{i-2}, \dots, \omega_i),$$

де

- $L -$ функція втрати (наприклад, перехрестна ентропія),
- $N -$ кількість слів у корпусі для тренування,
- $p(\omega_i | \omega_{i-1}, \omega_{i-2}, \dots, \omega_i) -$ ймовірність правильного слова ω_i з умовою на всі попередні слова,
- $\omega_i -$ і-те слово у послідовності.

Після попереднього навчання модель може бути адаптована (файн-тюнінг) для конкретних задач NLP, таких як класифікація тексту, відповіді на питання, переклад та інше, з використанням значно меншого набору даних. Процес файн-тюнінгу зазвичай включає незначну модифікацію архітектури моделі (наприклад, додавання специфічних для задачі виходів) та додаткове навчання, щоб адаптувати загальне розуміння мови, отримане під час попереднього навчання, до конкретних вимог завдання.

1.2.2 Encoder та Decoder

Архітектура трансформера складається з двох основних компонентів: енкодера і декодера.

Енкодера зчитує і аналізує вхідний текст і перетворює його в проміжне векторне представлення, яке містить у собі кодовану інформацію про весь текст.

Декодер використовує це представлення для генерації вихідного тексту, крок за кроком вирішуючи, яке слово буде наступним.

Обидва компоненти використовують блоки уваги та повнозв'язні (fully-connected) шари, але у декодері є додатковий блок уваги, який допомагає зосереджуватися на відповідних частинах вхідного тексту під час генерації відповіді. Схему роботи компонентів можна побачити на рисунку 1.

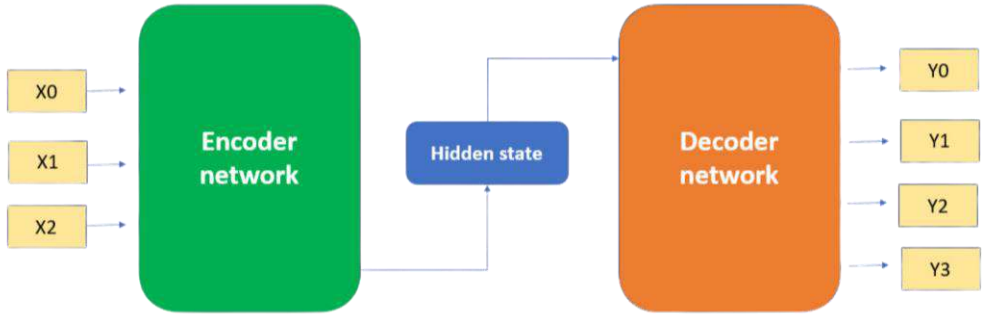

Рисунок 1. Схема роботи Encoder та Decoder.

1.2.3 Механізм Уваги

Механізм уваги в архітектурі трансформерів є ключовим елементом [1], який дозволяє моделі зосереджуватися на релевантних частинах вхідних даних для виконання конкретних завдань. Основна ідея полягає в тому, щоб визначити, наскільки кожен елемент послідовності (наприклад, кожне слово у реченні) є важливим для кожного іншого елемента в цій же послідовності. Для цього використовуються три ключові компоненти: вектори запитів (Q), ключів (K) та значень (V).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

де

- Q – матриця запитів (queries),
- K – матриця ключів (keys),
- V – матриця значень (values),

- d_k – розмірність ключів, яка використовується для масштабування скалярного добутку, з метою уникнення занадто великих значень перед активацією $softmax$, що може призвести до проблем з градієнтами.

Основа механізму уваги — це скалярний добуток між матрицею запитів Q та транспонованою матрицею ключів K^T. Цей добуток дає матрицю ваг уваги, де кожен елемент вказує на міру "важливості" або "уваги", яку один елемент (запит) повинен приділити кожному іншому елементу (ключу).

Ділення скалярного добутку на $\sqrt{d_k}$ зменшує варіації великих значень скалярного добутку, які можуть ускладнити навчання через зменшення градієнтів. Це допомагає підтримувати стабільність градієнтів, особливо для великих розмірів ключів.

Застосування функції $softmax$ до результату дозволяє перетворити скалярні значення в матрицю ваг уваги на інтервал [0, 1], де кожен рядок сумується до 1. Це перетворює ваги уваги в ймовірності, вказуючи на те, скільки "уваги" кожен запит повинен "платити" кожному ключу.

Нарешті, отримана після $softmax$ матриця ваг уваги множиться на матрицю значень V. Це визначає, як саме кожне значення буде враховуватись при виході, враховуючи ваги, визначені на основі запитів і ключів.

Отже формула механізму уваги дозволяє кожному елементу вихідної послідовності "звертати увагу" на всі елементи вхідної послідовності, що забезпечує динамічне врахування контексту. Механізм уваги є фундаментом для багатьох сучасних архітектур NLP і дозволяє ефективно обробляти довгі послідовності даних, виокремлюючи важливу інформацію та зв'язки між елементами.

1.2.4 Генерація відповідей

Етап генерації відповідей у GPT (Generative Pre-trained Transformer) базується на послідовному виборі слів на основі розподілу ймовірностей, що

залежить від контексту. Цей процес може бути представлений у вигляді ймовірностей для кожного можливого наступного слова в послідовності, даної попереднім контекстом.

Для кожного кроку генерації тексту, GPT оцінює розподіл ймовірностей наступного слова на основі контексту, який вже згенеровано. Ймовірність вибору наступного слова ω_i з урахуванням попереднього контексту $\omega_{<i}$ може бути визначена за наступною формулою:

$$P = (\omega_i|\omega_{<i}) = \frac{\exp(h_{i-1}^T e(\omega_i))}{\sum_{w'} \exp(h_{i-1}^T e(\omega'))},$$

де

- $P = (\omega_i|\omega_{<i}) -$ це ймовірність слова ω_i з урахуванням усього попереднього контексту $\omega_{<i}$,
- h_{i-1}– це вектор стану, який представляє контекст до моменту вибору слова ω_i отриманий з трансформера,
- $e(\omega_i) -$ це ембедінг слова ω_i,
- Сума в знаменнику пробігає по всіх можливих словах ω' у словнику, забезпечуючи нормалізацію ймовірностей.

Далі використовується функція $softmax$ для перетворення скалярних добутків у ймовірності, забезпечуючи, що сума ймовірностей усіх можливих наступних слів дорівнює 1.

Під час генерації тексту, можуть бути використані різні стратегії для вибору слів [1]:

- Жадібний Вибір: Вибір слова з найвищою ймовірністю на кожному кроці.
- Вибір з Оманливістю (Top-k sampling): Вибір слова випадково з топ-k найбільш ймовірних слів.

- Nucleus Sampling: Вибір випадкового слова з "ядра" розподілу ймовірностей, де вибираються слова, сумарна ймовірність яких досягає певного порогу P.

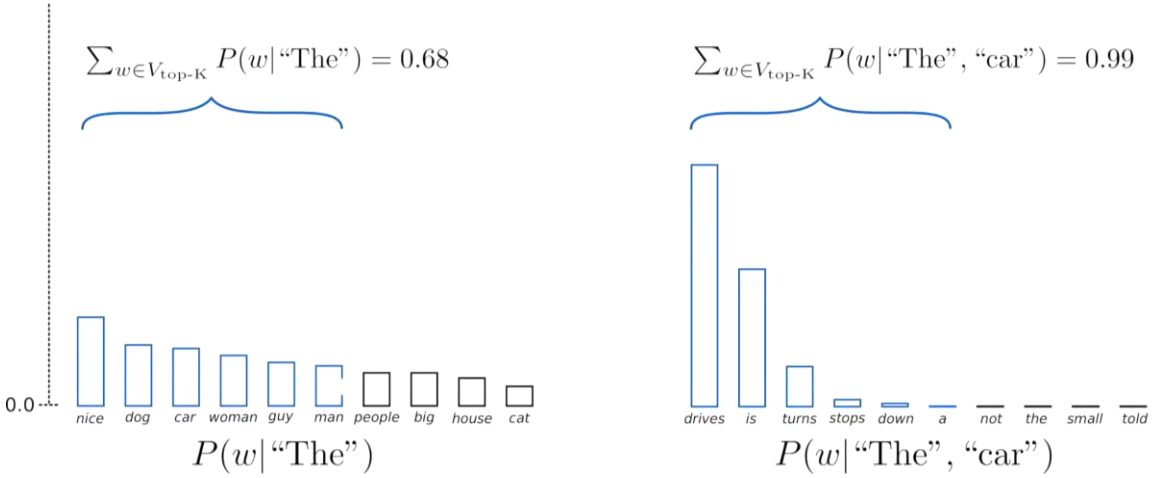

Рисунок 2. Ілюстрація ймовірностей слів у послідовності для вибору стратегії відбору слова.

Процес генерації відповідей у GPT - це механізм, який використовує знання, отримане під час попереднього навчання, для створення змістовного та контекстуально відповідного тексту. Використовуючи ітеративний підхід для вибору кожного наступного слова, GPT може генерувати тексти, що вражають своєю здатністю до збереження змістовності та логічності на протязі довгих послідовностей.

2. Інтеграція ШІ ChatGPT у середовище програмування SAS

SAS 9.4 (SAS Institute Inc., Кері, Північна Кароліна, США) — це керований командами пакет статистичних програм, який широко використовується для аналізу та візуалізації статистичних даних в сферах клінічних випробувань, банківських справах, диспетчерських службах.

Ефективна інтеграція ChatGPT у середовище програмування SAS може створити цінну можливість для підвищення продуктивності та досвіду програмування. Існує кілька практичних підходів, які програмісти SAS можуть вивчати для включення можливостей генерації мови ChatGPT у свій робочий процес [3]:

- Чат-інтерфейс у середовищі IDE: Один із ефективних методів полягає у інтеграції чат-інтерфейсу безпосередньо у інтегроване середовище розробки (IDE) SAS. Це дозволяє програмістам взаємодіяти з ChatGPT у реальному часі під час написання коду. Ця інтерактивна функція чату дозволяє програмістам задавати питання, шукати пропозиції коду або запитувати допомогу щодо конкретних завдань кодування, не переходячи між різними інструментами.
- Інтеграція на основі API: Інший підхід полягає в інтеграції ChatGPT через API (інтерфейси програмування застосунків). Це реалізує виклики API до ChatGPT з середовища програмування SAS. Програмісти можуть вводити свої запити або кодування та отримувати згенеровані частини коду або пояснення. Цей підхід пропонує гнучкість, дозволяючи програмістам інтегрувати ChatGPT у різні інструменти та інтерфейси SAS.
- Спеціалізовані плагіни та розширення: Розробники можуть створювати спеціалізовані плагіни або розширення для інструментів програмування SAS, що включають можливості ChatGPT. Ці плагіни можуть пропонувати функції, такі як автозавершення коду, адаптоване до контексту, автоматичне генерування коду на основі вимог або навіть

пояснення для секцій коду в реальному часі. Такі спеціалізовані розширення покращують середовище програмування SAS, надаючи підтримку безпосередньо в межах IDE.

- Керований робочий процес програмування: ChatGPT може слугувати провідником у процесі програмування. Програмісти можуть окреслювати свою логіку або потреби в кодуванні, а ChatGPT може генерувати відповідний код, пропонувати кроки оптимізації або навіть виявляти потенційні помилки. Цей підхід може допомогти як досвідченим, так і менш досвідченим програмістам у створенні ефективного коду SAS.
- Контекстуальна документація та допомога: Інтеграція ChatGPT для контекстуальної документації та допомоги також є цінною. Коли програмісти стикаються з незнайомими функціями або синтаксисом, вони можуть шукати пояснення, приклади або пропозиції кращих практик від ChatGPT. Це допомагає заповнити прогалини в знаннях і надає негайну допомогу у середовищі програмування.
- Автоматизація генерації завдань: Програмісти SAS можуть автоматизувати генерацію поширених завдань програмування. Наприклад, програмісти можуть описувати бажаний аналіз або результат, а ChatGPT може виробляти необхідний код. Цей підхід прискорює процес програмування, мінімізуючи зусилля з ручного кодування.
- Інтерактивна допомога при відлагодженні коду: Програмісти можуть описувати помилки, а ChatGPT може надавати роз'яснення щодо потенційних причин та рішень. Цей підхід оптимізує процес налагодження, скорочуючи час налагодження.

Вибір найбільш підходящого підходу до інтеграції залежить від таких факторів, як бажаний рівень взаємодії, глибина допомоги, вартість інтеграції та сумісність з існуючими інструментами SAS. Незалежно від використаного

підходу, кінцева мета залишається однаковою: використання можливостей ChatGPT для ефективної генерації коду, пропозицій та допомоги у вирішенні проблем у сфері клінічних випробувань.

2.1 З’єднання ChatGPT через SAS 9.4 IDE

Для того щоб інтегрувати ChatGPT у SAS 9.4 IDE, треба створити акаунт на сайті platform.openai.com. Після налаштування вашого акаунта найважливішим кроком є отримання секретного ключа API, який налаштовує доступ до API. Обов'язково скопіюйте та надійно збережіть цей ключ API для майбутнього використання.

У моїй програмі SAS ми будемо працювати з двома макро-змінними: &API_KEY і &QUESTION. Макро-змінна &API_KEY міститиме ваш секретний ключ API, тоді як макро-змінна &QUESTION слугуватиме як питання або пошуковий запит, який буде передаватись до ChatGPT.

Наступний програмний код дозволяє створити з’єднання з ChatGPT через створений API-ключ, відправити ваш запит з питанням, та отримати відповідь, який буде виведено напряму у вашу середу програмування:

```
      %macro RunGPTquery(query =);
            %let API_KEY= sk-XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX;
            %let QUESTION = %str(%"&query.%");

            /* Add model to request */
            filename in temp;
            data _null_;
            file in;
            put;
            put "{";
            put  '"model": "gpt-4", "messages": [{"role": "user", "content":
'"&question }]";
            put "}";
            run;
            /* Reference the file as in */
            filename resp "%sysfunc(getoption(WORK))/echo.json";
            proc http
             method="POST"
             url="https://api.openai.com/v1/chat/completions"
             ct="application/json"
             in=in
             out=resp;
             headers "Authorization" = "Bearer &api_key.";
            run;
            /* Parsing the response and output */
            libname response JSON fileref=resp;
```

```
            proc report data= response.choices_message;
            column content;
            define content / display "" style(column)=[cellwidth=6in fontsize=10pt
asis=ON];
            run;
      %mend RunGPTquery;
```

Наступний крок полягає у визові макро-коду з зовнішнього файлу та його виклику із запитом.

```
%include "/home/u43526594/SAS & ChatGPT/RunGPTquery.sas";

%RunGPTquery(query = SAS code to create boxplot by treatment group and save it
to .pdf format);
```

Після виконання коду результат буде згенерований з відповіддю на запит.

```
Here is an illustrative example of how you might generate the boxplot as request
ed:
```sas
ods graphics on;
ods pdf file="c:/my_folder/my_boxplot.pdf";
proc sgplot data=my_dataset;
   vbox my_variable / category=treatment_group;
   title "Boxplot by Treatment Group";
run;
ods pdf close;
ods graphics off;
```
```

Виконуючи код, запропонований ChatGPT, SAS створює боксплот та зберігає його за вказаним шляхом.

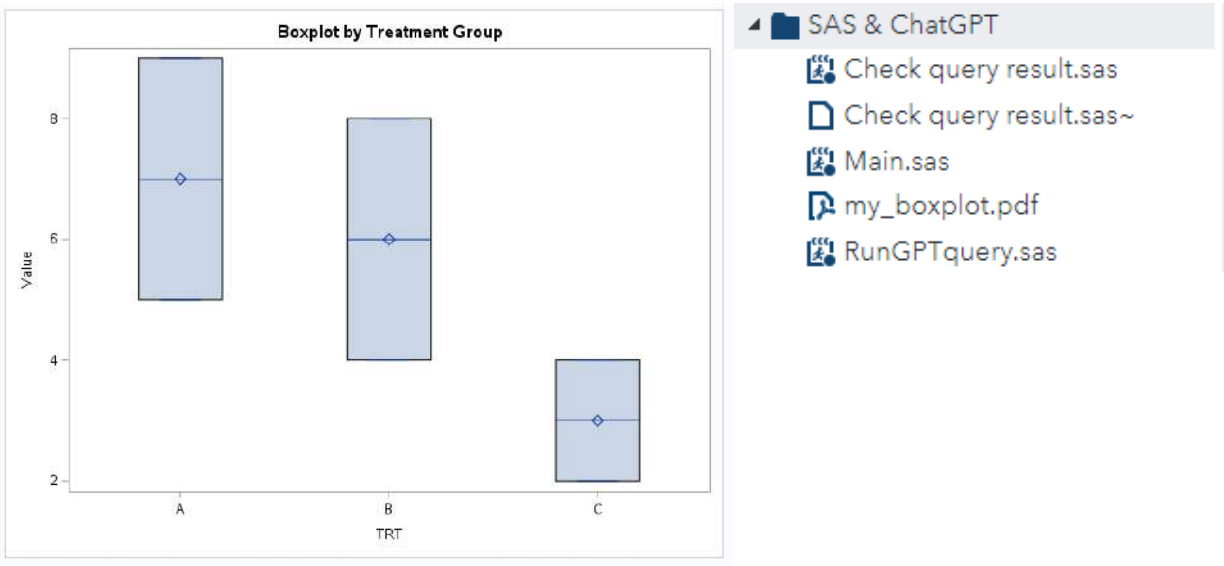

Рисунок 3. Результат виконання запропонованого ChatGPT кодом.

Інтеграція ChatGPT в середовище SAS через IDE пропонує ряд переваг, таких як допомога в реальному часі, ефективність та контекстуальна допомога. Однак, програмісти повинні бути обізнані щодо точності, потенційних

проблем з підключенням, обмежень у сфері застосування та питань конфіденційності даних.

Розуміючи ці переваги та недоліки, програмісти SAS можуть ухвалювати обґрунтовані рішення щодо того, коли і як ефективно використовувати можливості ChatGPT в контексті програмування клінічних випробувань.

3. Вирішення завдань за допомогою ШІ

3.1 Перевірка стандартів за допомогою ChatGPT

Стандарти CDISC (Clinical Data Interchange Standards Consortium) давно є ключовим елементом клінічних досліджень, забезпечуючи послідовність, ясність та впорядкованість обміну даними. Зокрема ADaM (Analysis Data Model) та SDTM (Standard Data Tabulation Model) є обов'язковими стандартами для представлення даних, статистичного аналізу та звітності у клінічних випробуваннях.

ChatGPT можна використовувати для генерації, перегляду та надання рекомендацій щодо створення наборів даних за стандартами SDTM та ADaM, забезпечуючи їх відповідність загальним стандартам CDISC. Потенціал такої інтеграції може радикально скоротити час, зусилля та потенційні помилки, пов'язані з ручною обробкою стандартних наборів даних.

Нижче ми розглянемо взаємодію між ChatGPT та програмуванням SAS, перевіряючи, чи присутні усі необхідні змінні в наборі даних SDTM.LB (Laboratory Dataset), згідно зі стандартом (дві необхідні змінні спеціально були видалені).

```
libname SDTM "/home/u43526594/SDTM";
proc contents data = SDTM.LB(drop = LBSEQ LBDTC)
 out = LB_VARS(keep = name label) noprint;
run;
data _null_;
        set LB_VARS end=last;
  length var_list $2000;
  retain var_list '';
        var_list = catx(' ', var_list, name);
        if last then call symputx('var_list', var_list);
run;
%put &var_list.;

%RunGPTquery(query = check if all the REQUIRED variables according to SDTM IG
v.3.3 are present in my LB dataset. The variables in my LB dataset are as follows:
&var_list.);
```

Нижче представлена відповідь моделі штучного інтелекту:

Comparing this list to your dataset, it seems you're missing the LBSEQ variable which is crucial for the order of observations. Also, LBDTC (Date/Time of Observation/Collection) is a required variable and needs to be included even though you have LBDY

Як ми можемо бачити, штучний інтелект виявив, що у нашому наборі даних не вистачає двох змінних, які, згідно зі стандартом, обов'язково повинні бути присутні.

3.2 Обробка програмного коду штучним інтелектом для створення документації

Створення специфікацій також є основною частиною щоденної рутини обробки клінічних даних, незалежно від того, чи йдеться про виробництво аналітичних даних (ADaM) чи моделі табуляції даних дослідження (SDTM). Це часто вимагає глибокої логіки та міцного розуміння мети та структури набору даних для створення змінних із даними пацієнтів.

Код SAS, що використовується для генерації цих змінних, може бути наданий у природній мові через ChatGPT. Це допомагає забезпечити, що як технічні, так і нетехнічні користувачі зрозуміють документацію специфікації.

Уявіть, що під час роботи зі специфікацією для набору даних з побічними ефектами пацієнтів (ADAE) ви зустрічаєте цей фрагмент коду і хочете створити інструкцію для змінної стартової дати побічного ефекту (ASTDY):

```
/*Unknown code*/
ASTDY=ASTDT-TRTSDT+(ASTDT>=TRTSDT);
/**/
%let code = %nrstr(ASTDY=ASTDT-TRTSDT+(ASTDT>=TRTSDT));


%RunGPTquery(query = Could you create derivation for my specification using this code
&code. and check it with IG compliance);
```

Результат виконання програмного коду:

```
ASTDY is derived based on the input variables ASTDT (start date of abnormal status)
and TRTSDT (start date of treatment).

If ASTDT is not missing and TRTSDT is not missing, ASTDY is the difference in days
between ASTDT and TRTSDT. If ASTDT is on or after TRTSDT, a day is added to that
difference.

Here is a step-by-step process, and its compliance with CDISC/CDASH standards for
clinical trials (IG = Implementation Guide).

1. First, check that ASTDT (start date of abnormal status) and TRTSDT (start date of
treatment) are not missing. According to CDISC/CDASH standards, dates and times should
be input in ISO 8601 format (YYYY-MM-DD or YYYY-MM-DDThh:mm:ss). The derivation cannot
be performed if either of these variables is missing.

2. Subtract TRTSDT from ASTDT to get the difference in the number of days.
```

3. Check if ASTDT is on or after TRTSDT. If it is, add one (+) to the difference in days. If it is not, don't add anything.

4. The result is ASTDY, the number of days from the start date of treatment to the start date of the abnormal status. If ASTDT is the same as or later than TRTSDT, it is counted as an additional day.

In terms of the IG compliance, this calculation is generally compliant with the CDISC/CDASH standards. The calculation is transparent, and the resulting derived variable (ASTDY) makes sense in the clinical context.

However, **it's always important to clearly define all variables and coding in a study protocol and Statistical Analysis Plan (SAP)** to ensure full transparency and compliance with the IG' recommendations. Performance of this calculation and its incorporation into analysis should be fully explained in the study documentation.

Тепер детально розглянемо результат який представив ChatGPT. Мета полягає в тому, щоб переконатися, що змінна ASTDY, яку ми отримали раніше, є правильно порахованою. Ми попросимо ChatGPT написати код SAS на основі встановлених правил, яка вивела модель. Ми пройдемо через цей зворотній процес, який використовується як перевірка якості, щоб забезпечити точність результатів.

```
%let rules = %nrstr(
If ASTDT is not missing and TRTSDT is not missing, ASTDY is the difference in days
between ASTDT and TRTSDT. If ASTDT is on or after TRTSDT, a day is added to that
difference.
Here is a step-by-step process
. . .
If ASTDT is the same as or later than TRTSDT, it is counted as an additional day.
);
%RunGPTquery(query = create SAS code to derive ASTDY variable using this rules:
&rules.);
```

Результат виконання програмного коду:

````
Following these instructions, SAS code that meets the CDISC/CDASH standards for
clinical trials could look like this:
```SAS
DATA new_dataset;
SET old_dataset;
*Check that ASTDT and TRTSDT are not missing;
IF NOT (MISSING(ASTDT) OR MISSING(TRTSDT)) THEN DO;
    *Calculate the difference between ASTDT and TRTSDT in days;
    ASTDY = ASTDT - TRTSDT;

    *Check if ASTDT is on or after TRTSDT. If it is, add one to the difference in
days;
    IF ASTDT >= TRTSDT THEN ASTDY = ASTDY +1;
END;
RUN;
```
````

Точний код був успішно згенерований за допомогою ChatGPT. Крім того, він перевірив наявність відсутніх даних і додав до ASTDY один день, якщо дата початку прийому першої дози досліджуваного препарату (TRTSDT)

відбулася після дати початку побічного ефекту (ASTDT). Більше того, ChatGPT додав коментарі в наданий код.

Ця секція оцінила здатність ChatGPT підтримувати програмування SAS у клінічних випробуваннях. Ми продемонстрували, як він може створювати нові змінні, аналізувати дані, щоб переконатися у їх відповідності стандартам SDTM та ADaM, а також писати код на основі специфікацій дослідження. Ці функції підсилюють нашу роботу та можуть покращити якість даних, які ми використовуємо у клінічних випробуваннях.

Але важливо пам'ятати, що ChatGPT був створений для допомоги людям, а не для самостійного виконання завдань або взаємодії з машинамию. Хоча ChatGPT і корисний, ви завжди повинні бути обережними з його виводом. Як і будь-який інструмент, може існувати шанс помилки. Тому ви завжди повинні перевіряти його роботу з офіційними документами дослідження та настановами.

Усі результати виводу потрібно ретельно перевіряти, щоб впевнитися, що вони відповідають строгим клінічним вимогам і стандартам CDISC.

4. Генерація тестових даних за допомогою моделі ШІ ChatGPT

Використання ChatGPT у розробці наборів даних клінічних випробувань вимагає точного та детального плану, який описує, як підготувати вхідні документи та організувати бажані результати. Недостатньо просто перетворити документи клінічних випробувань у дані, щоб зробити схему працездатною. Процедура також повинна дозволити штучному інтелекту розуміти та аналізувати складні статистичні та медичні дані, щоб автоматизувати точне та ефективне створення наборів даних у мові програмування SAS.

Після успішної інтеграції ChatGPT в середовище SAS, наступним кроком буде використання макрокоду для з'єднання з ChatGPT та відправлення запитів для визначення точного типу даних, які генеруються на виході. У нашому прикладі генеруються тестові дані про вимірювання життєвих ознак (Vital Signs) пацієнтів під час різних візитів. Для генерації синтетичних, але реалістичних даних життєвих ознак ми включаємо такі фактори: назви змінних, довжину та типи. Також важливо зазначити структуру бажаного набору даних та вміст кожної змінної.

```
%RunGPTquery(query = Create raw data set with sample data for Vital Signs
clinical trial which will represent some sample data. The structure should
be one row per subject per visit.
SubjectID should be character and represent unique number of patient up to
6 digits. VisitNumber should be numeric and represent number of visit for
each measurement subject visit.
Weight should be character and contain units and should not be too
different from visit to visit for one patient.
Height should be character with units and could be measured only once per
patient. BloodPressure should be character and contain systolic and
diastolic blood pressure with units.
Generate 20 records for 5 patients for different visits. The data set
should have 5 variables separated with comma each);
```

SubjectID, VisitNumber, Weight, Height, BloodPressure
P001, 1, 75kg, 170cm, 120/80 mmHg
P001, 2, 75.5kg, , 121/82 mmHg
P001, 3, 76kg, , 122/81 mmHg
P001, 4, 76kg, , 120/79 mmHg
P001, 5, 76.5kg, , 119/79 mmHg
P002, 1, 80kg, 180cm, 125/85 mmHg
P002, 2, 80.5kg, , 126/86 mmHg
P002, 3, 80kg, , 125/87 mmHg
P002, 4, 80.5kg, , 124/88 mmHg
P002, 5, 81kg, , 123/86 mmHg
P003, 1, 70kg, 165cm, 110/70 mmHg
P003, 2, 70.5kg, , 111/72 mmHg
P003, 3, 70kg, , 110/71 mmHg
P003, 4, 70.5kg, , 109/70 mmHg
P003, 5, 71kg, , 108/69 mmHg
P004, 1, 85kg, 175cm, 130/90 mmHg
P004, 2, 85.5kg, , 131/91 mmHg
P004, 3, 85kg, , 132/90 mmHg
P004, 4, 85.5kg, , 130/91 mmHg
P004, 5, 86kg, , 129/90 mmHg
P005, 1, 65kg, 160cm, 105/65 mmHg
P005, 2, 65.5kg, , 104/66 mmHg
P005, 3, 66kg, , 106/65 mmHg
P005, 4, 66.5kg, , 104/66 mmHg
P005, 5, 67kg, , 105/65 mmHg

Рисунок 4. Згенеровані ШІ дані.

Наступний крок полягає у збереженні створеного набору даних для подальшого дослідження. Щоб забезпечити цілісність створених синтетичних даних, збережемо вихідні дані у текстовий файл. Після створення текстового файлу використовується оператор infile SAS. Оператор infile написаний для зачитування текстових файлів і ретельної обробки вмісту, щоб гарантувати, що кожна змінна була правильно розпізнана. Процес призводить до створення набору даних у оригінальному форматі SAS, файлу .sas7bdat, що гарантує сумісність з багатьма аналітичними процедурами SAS і забезпечує легкий доступ та маніпулювання даними у SAS.

```
data my_dataset;
    infile '/home/u43526594/SAS & ChatGPT/sampledata.txt' dlm=',' dsd;
    length BloodPressure $12;
    input SubjectID $ VisitNumber Weight $ Height $ BloodPressure $;
run;
```

Columns
Select all
SubjectID
VisitNumber
BloodPressure
Weight
Height

Total rows: 25 Total columns: 5 Rows 1

	SubjectID	VisitNumber	BloodPressure	Weight	Height
1	P001	1	120/80 mmHg	75kg	170cm
2	P001	2	121/82 mmHg	75.5kg	
3	P001	3	122/81 mmHg	76kg	
4	P001	4	120/79 mmHg	76kg	
5	P001	5	119/79 mmHg	76.5kg	

Рисунок 5. Згенерований набір тестових даних у форматі .SAS7BDAT

4.1 Методи налаштування даних під час генерації

Наступний підхід полягає не лише у створенні даних; він включає точне моделювання сценаріїв клінічних випробувань, що відбуваються в реальному житті за допомогою штучних даних. Метою є створення набору даних, який відповідає всім вимогам, водночас відображаючи непередбачуваність реальних клінічних даних. Це можливо лише за допомогою ретельного підходу до моделювання даних, який враховує статистичні розподіли, взаємозв'язки між змінними та властиві шаблони, знайдені у клінічних випробуваннях.

Ми прагнемо підвищити точність прогнозних моделей у практичних застосуваннях, включаючи ці фактори у створення зразкових наборів даних, які забезпечують основу для тренування.

Наступні фактори та дії можуть бути вжиті для поліпшення або навмисного погіршення якості генерованих даних для наших цілей.

Таблиця 1 Методи для точного налаштування згенерованих даних

Тип аналізу	Обрані дії
РОЗПОДІЛ ДАНИХ ТА РЕПРЕЗЕНТАТИВНІСТЬ	
Дослідіть, як розподіл генерованих даних може бути порівняний з передбачуваним розподілом, отриманим з минулих даних або досвіду у відповідній галузі. Аналізуйте частотні підрахунки для категоріальних змінних, а також середнє, медіану, варіацію та діапазон для неперервних змінних.	Використовуйте статистичні тести (наприклад, тест Колмогорова-Смірнова для неперервних даних) для порівняння розподілів та забезпечення репрезентативності генерованих даних відносно цільової популяції.
ПАТТЕРНИ ВІДСУТНІХ ДАНИХ	
Генеруйте моделі відсутніх даних у зразкових даних. Відсутність може бути встановлена як випадкова або мати шаблон, що може спотворити аналіз.	Впроваджуйте механізми відсутніх даних (наприклад, MCAR, MAR, NMAR) у процес генерації, щоб точно відображати моделі відсутніх даних, спостережувані в реальних клінічних випробуваннях.

КОРЕЛЯЦІЯ ТА ПРИЧИННІСТЬ	
Створіть винятки або екстремальні значення у генерованих даних, визначаючи, чи є вони реалістичними або артефактами процесу генерації.	Визначте правила або обмеження у процесі генерації даних для обробки винятків, забезпечуючи їх узгодженість з реалістичними клінічними сценаріями..
ДОТРИМАННЯ ПЛАНУ СТАТИСТИЧНОГО АНАЛІЗУ (SAP)	
Переконайтеся, що генеровані дані відповідають специфікаціям, викладеним у SAP, включаючи аналіз популяції, статистичні методи та визначення фінальних цілей.	Регулюйте процес генерації, щоб дані відповідали (або не відповідали) критеріям аналізу, зазначеним у SAP.
ЗОВНІШНЯ ВАЛІДАЦІЯ	
Де це можливо, проведіть валідацію генерованих даних проти зовнішніх наборів даних або критеріїв, щоб перевірити їх точність і адекватність.	Регулюйте процес генерації даних на основі результатів зовнішньої валідації, забезпечуючи узгодженість генерованих даних зі зовнішніми стандартами та очікуваннями.

4.2 Застосування моделі генерації даних у клінічних випробуваннях

Цей розділ досліджує реальне використання ChatGPT у інтеграції з мовою SAS в контексті клінічних випробувань, надаючи прикладний випадок, який демонструє, як ChatGPT може бути використаний для поліпшення наборів даних SAS, що генеруються з документів дослідження ранньої фази клінічного випробування. У цій роботі буде зосереджено увагу на описі складного процесу забезпечення дотримання ChatGPT детальних специфікацій, перелічених у протоколі дослідження, плані статистичного аналізу (SAP) та формах звітності за випадками (CRF). Ця інтеграція не лише показує, наскільки добре ChatGPT може аналізувати і розуміти складні клінічні дані, але й підкреслює модифікації та зміни, які необхідні для адаптації ChatGPT до конкретних вимог різних клінічних випробувань.

Програмний код наданий нижче дозволяє надсилати запити до ChatGPT, включаючи важливу інформацію, отриману з SAP (Плану статистичного аналізу) та CRF (Форми звітності). Він включає детальний список змінних, їх відповідні значення та розгорнуті інструкції для їх генерації. Мета полягає у тому, щоб показати, як ChatGPT може автоматизувати перетворення неструктурованих, неаналізованих «сирих» клінічних даних у структуровані набори даних, значно спрощуючи процес генерації даних.

Для прикладу візьмемо будь-який протокол дослідження та план статистичного аналізу (SAP) на сайті clinicaltrials.gov, які є публічно доступними. У нашому прикладі аналізуються протокол дослідження та SAP дослідження «Терапевтичний потенціал імбиру в лікуванні астми (GINGER)» з метою створення набору «сирих» даних медичної історії для конкретного клінічного дослідження.

4.1 Inclusion Criteria

Subjects must meet all of the inclusion criteria to participate in this study.

1. Adults aged 18 years and above, with mild to severe persistent asthma will be enrolled as defined by The NIH National Asthma Education and Prevention Program (NAEPP) Guidelines for Assessing Asthma Severity[1] (please see table page 9). Subjects will have asthma which is not optimally controlled as defined by Asthma Control Test (ACT) score, despite the current use of inhaled corticosteroids with or without inhaled long-acting beta agonists.
2. Treatment with inhaled corticosteroids (ICS) +/- long acting β-agonists/long acting muscarinics and montelukast;

4.2 Exclusion Criteria

Subjects with any of the exclusion criteria will be excluded from study participation.

1. Other major chronic illnesses: Conditions which in the judgment of the study physician would interfere with participation in the study, e.g., non-skin cancer, uncontrolled diabetes mellitus, coronary artery disease, congestive heart failure, stroke, severe hypertension, renal failure, liver disorders, malabsorption disorders, immunodeficiency states, major neuropsychiatric disorder;
2. Cardiovascular problems:
    a. Myocardial infarction or stroke in last 3 months
    b. Uncontrolled hypertension

Рисунок 6. Приклад критеріїв відбору пацієнтів для обраного дослідження, скопійований з плану статистичного аналізу.

PREVIOUS MEDICAL HISTORY

Is there any relevant medical history in the following systems?

Code	System	*Yes	No	Code	System	*Yes	No
1	Cardiovascular			9	Neoplasia		
2	Respiratory			10	Neurological		
3	Hepato-biliary			11	Psychological		
4	Gastro-intestinal			12	Immunological		
5	Genito-urinary			13	Dermatological		
6	Endocrine			14	Allergies		
7	Hematological			15	Eyes, ear, nose, throat		
8	Musculo-skeletal			00	Other		

*If ***YES*** for any of the above, enter the code for each condition in the boxes below, give further details (including dates) and state if the condition is currently or potentially active. If giving details of surgery, please specify the underlying cause. Use a separate line for each condition.

		Currently Active?	
Code	**Details (including dates)**	**Yes**	**No**

Рисунок 7. Приклад CRF форми, яку заповнює лікар про медичну історію пацієнта.

Спочатку ми створимо дві макрозмінні, які міститимуть необхідну інформацію про дослідження. Макрозмінна **SAP_INSTRUCTIONS** міститиме критерії відбору, а макрозмінна **CRF_page** данні шаблон форми, яку заповнює лікар.

```
%let SAP_INSTRUCTIONS = %nrstr(
    Medical History: Relevant medical history, including history of current
disease, other pertinent respiratory history, and information regarding
underlying diseases will be recorded at the Screening visit and confirmed
during the randomization visit. Inclusion Criteria
Subjects must meet all of the inclusion criteria to participate in this
study.
. . .
2. Cardiovascular problems:
a. Myocardial infarction or stroke in last 3 months
b. Uncontrolled hypertension
c. Known aortic aneurysm
. . .
);

%let CRF_page = %nrstr(
    PREVIOUS MEDICAL HISTORY
Is there any relevant medical history in the following systems?
Code      System              *Yes      No                Code System
    *Yes  No
1   Cardiovascular                      9       Neoplasia
2   Respiratory                         10      Neurological
3   Hepato-biliary                      11    Psychological
4   Gastro-intestinal                   12    Immunological
5   Genito-urinary                      13    Dermatological
6   Endocrine                            14   Allergies
7   Hematological                             15    Eyes, ear, nose, throat

8
. . .
    Currently Active?
Code      Details (including dates)     Yes   No

);
```

Після створення макрозмінних ми можемо надіслати наш запит із викликом наших макрозмінних у ньому.

```
%RunGPTquery(query = Create raw Medical History data set in clinical trial
which will represent some sample data. The structure should be one row per
subject per medical term per timepoint.
SubjectID should be character and represent unique number of patient. Date
of collection should be different per each subject.
Medical History disorder description should be character. Start and end
date should be separated by day/month/year variables each.
Ongoing variable should be 1 or missing. Comment variable should be
character and can be missing for the most of medical disorders.
The data set should consist of 11 variables separated with comma each.
create at least 5 subjects and 10 disorders per each.
Some subjects can have exclusion criteria. Here are SAP details
&SAP_INSTRUCTIONS. Here is CRF page &CRF_page.);
```

Результат виконання цього запита має наступну форму:

Total rows: 25 Total columns: 11 Rows 1-25

SubjectID	CollectionDate	DisorderDescription	Start...	StartD...	StartDat...	End...	End...	EndDat...	Ongoi...	Comment
103	03-03-20	Asthma	01	03	2015				1	Controlled with medication
104	04-03-20	Arthritis	01	01	2019				1	Treated with drugs
103	03-03-20	Mood disorders	02	10	2018				1	Under therapy
102	02-03-20	Anxiety	03	01	2018				1	Under therapy
104	04-03-20	Bronchitis	03	03	2017				1	Controlled with medication
102	02-03-20	Diabetes	04	05	2017				1	Controlled with medication
104	04-03-20	Hypercholesterolemia	05	05	2017	05	05	2018		Lifestyle changes
103	03-03-20	Diabetes	05	09	2017				1	Controlled with diet
105	05-03-20	Depression	06	08	2018				1	Treated with therapy and medication
105	05-03-20	Diabetes	08	06	2016				1	Controlled with medication
101	01-03-20	Obesity	08	07	2014	14	07	2018		Lost weight
101	01-03-20	Hypercholesterolemia	09	11	2015	09	11	2016		Lifestyle changes
105	05-03-20	Asthma	10	04	2015				1	Controlled with medication
101	01-03-20	Diabetes	10	09	2017				1	Controlled with diet
104	04-03-20	Osteoporosis	10	02	2016				1	Treatment ongoing
102	02-03-20	Depression	12	02	2018				1	Controlled with medication
101	01-03-20	Asthma	12	02	2015	29	02	2020	1	Controlled with medication
103	03-03-20	Lower back pain	14	12	2019				1	Physical therapy ongoing
102	02-03-20	Asthma	14	04	2016				1	Controlled with medication
103	03-03-20	Cardiovascular diseases	15	08	2016				1	Treated with drugs & lifestyle changes
105	05-03-20	Hypertension	15	07	2017				1	Treated with lifestyle changes
101	01-03-20	Hypertension	15	06	2016				1	Controlled with medication
105	05-03-20	Arthritis	20	09	2019				1	Physical therapy ongoing

Рисунок 8. Згенерований набір даних збережений у SAS7BDAT форматі.

Як ми бачимо, ChatGPT успішно створив тестовий набір даних, проаналізувавши критерії включення/виключення. Він не тільки створив дані про захворювання на астму для кожного суб'єкта (що логічно було включити), але й залишив цю хворобу невилікуваною майже для кожного суб'єкта.

Збільшення кількості існуючих записів — це ще один метод генерації даних за допомогою штучного інтелекту. Перший крок полягає в створенні двох макрозмінних. У першу макрозмінну будуть включені всі змінні та їх метадані, які зараз включені в сирий набір даних лабораторних тестів. Безпосередньо дані, пов’язані з цими змінними, будуть у другій макрозмінній. Програмний код нижче, має на меті точно прочитати ці дані та зберегти їх у макрозмінних, розділяючи їх за допомогою спеціальних символів. Далі нам потрібно розв’язати макрозмінні у параметрі макрофункції %RunGPTquery, яка включає запит на створення набору даних.

```
%include "/home/u43526594/SAS & ChatGPT/RunGPTquery.sas";
libname raw "/home/u43526594/rawdata";

proc contents data = raw.labtest
                out = lab_vars(keep = name label) noprint;
run;
```

```
data _null_;
    set lab_vars end=last;
    length var_list var_labels $2000;
    retain var_list var_labels '';

    var_labels = cats(var_labels, name,'(',label,')','|');
    var_list = catx(',', var_list, name);

    if last then do;
          call symputx('var_list', var_list);
          call symputx('var_labels', var_labels);
    end;
run;


/*to keep the order of vars as per proc contets*/
proc sql;
    create table labtest as
          select &var_list from raw.labtest;
quit;
data labtestnn;
    set labtest;
    array vars _character_ ;
    do over vars;
          if missing(vars) then vars = "|";
    end;
run;

data _null_;
    set labtestnn;
    length lab_values $2000;
    retain lab_values '';

    /*take only one record*/
    if _N_ = 1 then do;
          lab_values = catx('|',lab_values, &var_list.);
          call symputx('lab_values', substr(lab_values,2));
    end;
run;

%RunGPTquery(query = Create raw data set consisted of laboratory test
results in clinical trial which will represent some sample data.
The structure should be one row per subject per labtest per visit.
Create the following variables: &var_labels..
Here is data from the first row: &lab_values..
Create 5 more data rows based on the data above.
Create some outrange data);
```

Нижче представлена відповідь штучного інтелекту, яка була збережена в txt формат, а згодом транспонована у SAS7BDAT формат.

ABNORM	CLIN_SIGN	COL_DT	COMMENT	NOT_DONE	OUT_RNGE	PAT_NUM	RHI_SI	RLO_SI
		2017-12-10T11:33:14				001	147	44
		2017-12-12T11:33:15			Yes	002	147	44
	Yes	2017-12-14T11:33:16				003	147	44
		2017-12-15T11:33:17		Yes		004	147	44
	Yes	2017-12-16T11:33:18				005	147	44
		2017-12-18T11:33:19				006	147	44

RSLT_SI	SITE	STUDY	T_GRP	T_NAME	T_NAMECD	UNIT_SI	VIS
144	001	AAA-BBB-CCC	HEMA	Eosinophils WBC absolute counts	EOS	10^9/L	Follow-up
150	001	AAA-BBB-CCC	HEMA	Eosinophils WBC absolute counts	EOS	10^9/L	Follow-up
160	001	AAA-BBB-CCC	HEMA	Eosinophils WBC absolute counts	EOS	10^9/L	Follow-up
	001	AAA-BBB-CCC	HEMA	Eosinophils WBC absolute counts	EOS	10^9/L	Follow-up
130	001	AAA-BBB-CCC	HEMA	Eosinophils WBC absolute counts	EOS	10^9/L	Follow-up
140	001	AAA-BBB-CCC	HEMA	Eosinophils WBC absolute counts	EOS	10^9/L	Follow-up

Рисунок 9. Відображення згенерованих даних з вхідних параметрів та запиту.

ChatGPT зміг не тільки згенерувати критичні значення, які перевищують допустимі межі, але й правильно позначити їх як клінічно значущі. Це важливе зауваження підкреслює, наскільки добре модель здатна створювати штучні дані, які виглядають реальними. Значущим є те, що однією з найбільших переваг створених даних буде їх гнучкість щодо параметрів, переданих у аргумент нашої функції запиту.

Ця динаміка показує корисність штучного інтелекту для розробки тестових наборів даних, які можна змінювати за запитом. Використовуючи наш код, ми надаємо можливість мати динамічний варіант створення «сирого «набору даних відповідно до конкретних потреб у дослідженнях або експериментальних ситуаціях.

5. Врахування регуляторних обмежень та конфіденційності

ChatGPT та подібні технології штучного інтелекту в клінічних випробуваннях є справжніми можливостями для подальшого підвищення ефективності та інновацій в управлінні даними. Однак, ця інтеграція не відбувається без викликів та обмежень. Серед них є точність даних у згенерованих відповідях, упередження, які можуть виникнути, та обмеження програмного забезпечення в розумінні складних концепцій у медицині та статистиці. Це ключові та чутливі питання, тому що безпека пацієнтів є пріоритетом.

5.1 Виклики та обмеження

Деякі з цих викликів та обмежень, які потрібно врахувати, включають наступне:

Точність згенерованих даних.

1. Якість даних: Точність даних, згенерованих за допомогою ChatGPT, повністю залежить від якості та деталей наданого вводу. Неправдиві та неповні дані можуть виробляти неправдивий або неповний набір даних і додатково ризикувати цілісністю аналізу клінічних випробувань.
2. Неправильно інтерпретований контекст: Моделі ШІ можуть неправильно інтерпретувати складний клінічний сценарій або навіть статистичне точне налаштування, що призводить до неправильного представлення клінічних реальностей.
3. Потенційні упередження: Відповіді ChatGPT генеруються моделлю, навченою на великому текстовому наборі даних. Якщо під час навчання дані мають будь-які упередження щодо статі, етнічної приналежності, поширеності захворювань або результатів лікування, то такі упередження ймовірно відобразяться в згенерованих даних і вплинуть на представлені результати.

4. Клінічні знання: Незважаючи на одну з найкращих реалізацій ШІ для обробки природної мови, ChatGPT не зможе зрозуміти багато критичних медичних термінологій і клінічних протоколів з глибоким планом статистичного аналізу. Це введе упередження або помилки в згенерованих даних.

Етичні та регуляторні питання.

1. Конфіденційність пацієнтів: Конфіденційність даних пацієнтів залишається на вищому рівні. Використання мовних моделей ChatGPT слід дотримуватися правил захисту, таких як HIPAA у США, GDPR в Європі, та загалом відповідних місцевих законів, які можуть бути проблематичними при використанні даних пацієнтів.
2. Регуляторне схвалення: Чітке та прийнятне використання даних, згенерованих штучним інтелектом у клінічних випробуваннях для регуляторних органів, вважається першим кроком. Це ключові критерії, і це робить схвалення складним через новизну технології в цій галузі та додаткові проблеми, пов'язані з цілісністю даних та упередженням.

5.2 Обмеження регуляторних органів

Управління з харчових продуктів та лікарських засобів США (FDA), відзначає зростання використання AI/ML у всьому спектрі терапевтичних засобів та на всіх етапах розробки лікарських засобів завдяки їхній здатності модернізувати та покращити багато процесів, серед яких, зокрема, ті, що залучені до відкриття лікарських засобів, клінічних випробувань, після маркетингового спостереження та фармацевтичного виробництва. [4]

FDA відзначила збільшення кількості заявок на ліки та біологічні препарати, що при дослідженні використовують AI/ML, при цьому в 2021 році їх було подано 100. Насправді, позиція FDA щодо використання AI/ML у розробці ліків залишається пов'язаною з першою поправкою: переконатися, що нові ліки розробляються з основною увагою на їх безпеку та ефективність.

Усвідомлюючи, що впровадження AI/ML несе унікальний набір викликів з кожною новою можливістю, FDA активно намагається переосмислити власну регуляторну рамку в інноваційну, яка була б представлена як дружня та гнучка, приймаючи необхідні заходи для захисту громадського здоров'я. [4]

Залучення FDA до використання AI/ML у розробці ліків є збалансованим прагненням до інновацій з урахуванням безпеки пацієнтів. FDA запевняє, що регуляторне середовище продовжує підтримувати передові технології, одночасно гарантуючи встановлення шлях вперед для наступних поколінь практик розробки ліків.

Кодекс федеральних правил під назвою 21 (CFR Title 21) [5], який є значущим документом регулювання, адміністрований FDA, додає суттєві вимоги, які гарантують безпеку, ефективність та безпеку клінічних випробувань.

Нижче наведені додаткові правила, визначені цим документом:

- Збереження записів: Згідно з правилами, електронні записи повинні зберігатися таким чином, що доступ до них можливий в будь-який час протягом терміну зберігання записів. Інтеграції ШІ не повинні ускладнювати збереження та відновлення даних.
- Забезпечення повноти даних: Наголошується на важливості збереження повних даних, зібраних з усіх тестів, обстежень та аналізів, щоб гарантувати дотримання встановлених специфікацій і стандартів.
- Перегляд даних: Потрібно регулярно перевіряти дані на повноту, якість і узгодженість з прийнятими стандартами. Інструменти AI, такі як ChatGPT, повинні сприяти цьому процесу, а не перешкоджати йому.

Також існують інші правила, які завжди слід враховувати при інтеграції будь-яких нових інструментів у перспективу клінічних випробувань: Електронні записи, Електронні підписи, Захист прав людини, Інституційні оглядові

комісії, Регулювання якості, Виключення дослідницьких пристроїв, Фінансове розкриття інформації клінічними дослідниками та інші. [5]

ВИСНОВКИ

Архітектура GPT (Generative Pre-trained Transformer) ілюструє потужний злет в області обробки природної мови, втілюючи собою витончену здатність моделювати мовні залежності та контекст на основі масштабного попереднього навчання та файн-тюнінгу. Завдяки механізму уваги, GPT здатна фокусуватися на різних аспектах вхідних даних, значно покращуючи здатність розуміння та генерації тексту. Це досягається через інноваційне використання трансформерів, які дозволяють ефективно враховувати як близькі, так і далекі залежності в тексті. Перехресна ентропія, використовувана як функція втрати під час навчання, забезпечує здатність точно оцінювати відповідність між прогнозованими та реальними даними, оптимізуючи модель для широкого спектра завдань обробки тексту. У сукупності, ці елементи формують основу для створення дуже потужних та гнучких моделей машинного навчання, що можуть адаптуватися до різноманітних завдань обробки природної мови, відкриваючи нові можливості для досліджень та застосувань у цій галузі.

Використання ChatGPT у процедурах програмування SAS для клінічних випробувань може мати значний вплив. Наприклад, це може повністю змінити спосіб збору даних і скоротити час, витрачений на клінічне дослідження. Клінічні програмісти можуть більше зосередитися на проектуванні експерименту, інтерпретації даних та етичних питаннях, автоматизувавши деякі процеси обробки та аналізу даних. Це збільшення продуктивності є не просто зручністю; це також може підсилити розробку лікарських засобів та їх введення на ринок.

Завдяки автоматизації рутинних завдань, таких як попередній аналіз даних, валідація даних та підготовка звітів, генерація тестових даних, аналітики даних та програмісти можуть зосередитися на більш складних і важливих аспектах клінічних випробувань. Таке залучення інтелектуальних технологій дозволяє клінічним дослідженням бути більш гнучкими та

адаптивними до нових наукових знахідок і регуляторних вимог, зменшуючи тим самим час від розробки препарату до його виходу на ринок.

Важливо визнати, що незважаючи на великий потенціал, у ChatGPT є обмеження. Це включає занепокоєння щодо конфіденційності даних та нинішню малоздатність моделі адекватно розуміти специфіку клінічних випробувань. Відповідно, майбутні дослідження повинні зосередитися на покращенні розуміння моделлю статистичної та медичної мови та впровадженні сильніших заходів безпеки.

Розвиток здатності штучного інтелекту до глибшого розуміння медичних та статистичних термінологій може значно покращити його застосування в клінічних випробуваннях, зробивши ці інструменти більш надійними та корисними для медичних фахівців. Одночасно, необхідно забезпечити, що використання таких технологій не порушує правила конфіденційності та захисту даних, які критично важливі для збереження довіри пацієнтів та цілісності досліджень.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Tufts CSDD White Paper Fall 2022. “Benchmarking and optimizing the process for adopting innovations supporting clinical trial execution”, 2022. URL: csdd.tifts.edu/white-papers/benchmarking_and_optimizing_the_process_for_adopting_innovations.

2. Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Łukasz, Polosukhin Illia “Attention is All you Need”. 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

3. Skliar Illia. 2023. “Boosting SAS Programming Efficiency with ChatGPT: A Clinical Trials Perspective.” PHUSE EU Connect, Birmingham: PHUSE. URL: https://phuse.s3.eu-central-1.amazonaws.com/Archive/2023/Connect/EU/Birmingham/PAP_CM04.pdf.

4. U.S FOOD & DRUG ADMINISTRATION, “Using Artificial Intelligence & Machine Learning in the Development of Drug & Biological Products”. URL: https://www.fda.gov/media/167973/download.

5. FDA website. “CFR – Code of Federal Regulations Title 21”, 2023 URL: accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm.

6. OpenAI platform documentation. URL: platform.openai.com/docs/quickstart/build-your-application.