ChatGPT based comments on a scientific manuscript

Marian Klose

2025-01-13

# Library

# Load packages  
import math  
import nltk  
from nltk.tokenize import sent\_tokenize  
from openai import OpenAI  
import json  
from docx import Document  
from pydantic import BaseModel  
from datetime import datetime

# NLTK options

# download the punkt tokenizer  
nltk.download("punkt\_tab")

True

# OpenAI options

# Initialize OpenAI client  
client = OpenAI()  
  
# Define model  
gpt\_model = "gpt-4o"

# Functions

# Function to split text into n nearly equal chunks  
def split\_text\_into\_enumerated\_sentences(text):  
 sentences = sent\_tokenize(text)  
 enumerated\_sentences = {i + 1: sentence for i, sentence in enumerate(sentences)}  
 return enumerated\_sentences  
  
# Function to split the enumerated dictionary into n parts  
def split\_enumerated\_sentences\_into\_chunks(enumerated\_sentences, n\_chunks):  
 total\_sentences = len(enumerated\_sentences)  
 chunk\_size = math.ceil(total\_sentences / n\_chunks)  
 chunks = []  
 keys = list(enumerated\_sentences.keys())  
 for i in range(0, total\_sentences, chunk\_size):  
 chunk = {key: enumerated\_sentences[key] for key in keys[i:i + chunk\_size]}  
 chunks.append(chunk)  
 return chunks  
  
# Function to add a comment to a sentence in Word format (for quarto / pandoc)  
def add\_word\_comment(sentence, comment):  
 date = datetime.now().isoformat()  
 return f'[{comment}]{{.comment-start id="0" author="ChatGPT" date="{date}"}}{sentence}[]{{.comment-end id="0"}}'

# Read and process text

# Load the document  
doc\_path = "pharmacokinetics.docx"  
document = Document(doc\_path)  
  
# Extract full text from the docx  
full\_text = "\n".join([para.text for para in document.paragraphs])  
  
# Split the text into an enumerated dictionary of sentences  
enumerated\_sentences = split\_text\_into\_enumerated\_sentences(full\_text)  
  
# Display the result from element 0 to 5  
print(list(enumerated\_sentences.items())[0:5])  
  
# Split the enumerated sentences into n chunks  
chunks = split\_enumerated\_sentences\_into\_chunks(enumerated\_sentences, 2)

[(1, 'Pharmacokinetics\nPharmacokinetics (from Ancient Greek pharmakon "drug" and kinetikos "moving, putting in motion"; see chemical kinetics), sometimes abbreviated as PK, is a branch of pharmacology dedicated to describing how the body affects a specific substance after administration.'), (2, '[1] The substances of interest include any chemical xenobiotic such as pharmaceutical drugs, pesticides, food additives, cosmetics, etc.'), (3, 'It attempts to analyze chemical metabolism and to discover the fate of a chemical from the moment that it is administered up to the point at which it is completely eliminated from the body.'), (4, "Pharmacokinetics is based on mathematical modeling that places great emphasis on the relationship between drug plasma concentration and the time elapsed since the drug's administration."), (5, 'Pharmacokinetics is the study of how an organism affects the drug, whereas pharmacodynamics (PD) is the study of how the drug affects the organism.')]

# System and user messages

# Define system and base user messages for review  
system\_message = "You are a professor reviewing a scientific manuscript. You receive a text excerpt for review and each sentence of the text has an unique identifier. Your job is to provide a structured output in a JSON style to indicate for which identifier you have feedback and what the feedback is."  
base\_user\_message = (  
 "Please review the following text excerpt for style, errors, conciseness, and understandability. "  
 "Provide concise (!) comments for single sentences. Your answer should be short. "  
 "Provide structured feedback in JSON format indicating for which sentence you have which feedback."  
 "It is very important that you do not have to provide feedback for every sentence and you should only provide feedback if you have something to say."  
)

# Define JSON structure

# define the pydantic structure for a single sentence feedback  
class SingleSentenceFeedback(BaseModel):  
 sentence\_id: int  
 feedback: str  
  
# define the pydantic structure for a list of feedbacks  
class MultipleSentencesFeedback(BaseModel):  
 feedback\_list: list[SingleSentenceFeedback]

# Process each chunk

# Collect feedback for all chunks  
all\_feedback = []  
  
# Loop over each chunk, call the API, and process the response  
for idx, chunk in enumerate(chunks):  
 # print message to user  
 print(f"Processing chunk {idx + 1}/{len(chunks)}")  
  
 # Prepare user message for the current chunk as JSON  
 user\_message = f"{base\_user\_message}\n\nExcerpt:\n{json.dumps(chunk)}"  
  
 try:  
 # Make the API request  
 completion = client.beta.chat.completions.parse(  
 model=gpt\_model,  
 messages=[  
 {"role": "system", "content": system\_message},  
 {"role": "user", "content": user\_message}  
 ],  
 response\_format=MultipleSentencesFeedback  
 )  
  
 # Extract the parsed response   
 feedback = completion.choices[0].message.parsed  
  
 # Extend the all\_feedback object  
 all\_feedback.extend(feedback.feedback\_list)  
  
 except Exception as e:  
 print(f"Error at chunk {idx}: {e}")

Processing chunk 1/2  
Processing chunk 2/2

# Apply comments to the document

# Apply comments to the sentences  
for feedback\_item in all\_feedback:  
 sentence\_id = feedback\_item.sentence\_id  
 feedback\_text = feedback\_item.feedback  
 if sentence\_id in enumerated\_sentences:  
 original\_sentence = enumerated\_sentences[sentence\_id]  
 commented\_sentence = add\_word\_comment(original\_sentence, feedback\_text)  
 enumerated\_sentences[sentence\_id] = commented\_sentence  
  
# Combine sentences back into full text  
commented\_text = " ".join(enumerated\_sentences.values())

# Print document

# Print the commented text  
print(commented\_text)

Pharmacokinetics Pharmacokinetics (from Ancient Greek pharmakon “drug” and kinetikos “moving, putting in motion”; see chemical kinetics), sometimes abbreviated as PK, is a branch of pharmacology dedicated to describing how the body affects a specific substance after administration. [1] The substances of interest include any chemical xenobiotic such as pharmaceutical drugs, pesticides, food additives, cosmetics, etc. It attempts to analyze chemical metabolism and to discover the fate of a chemical from the moment that it is administered up to the point at which it is completely eliminated from the body. Pharmacokinetics is based on mathematical modeling that places great emphasis on the relationship between drug plasma concentration and the time elapsed since the drug’s administration. Pharmacokinetics is the study of how an organism affects the drug, whereas pharmacodynamics (PD) is the study of how the drug affects the organism. Both together influence dosing, benefit, and adverse effects, as seen in PK/PD models. ADME A number of phases occur once the drug enters into contact with the organism, these are described using the acronym ADME (or LADME if liberation is included as a separate step from absorption): Liberation – the process of the active ingredient separating from its pharmaceutical formulation. [3][4] See also IVIVC. Absorption – the process of a drug entering into systemic circulation from the site of administration Distribution – the dispersion or dissemination of substances throughout the fluids and tissues of the body. Metabolism (or biotransformation, or inactivation) – the chemical reactions of the drug and irreversible breakdown into metabolites (e.g. by metabolic enzymes such as cytochrome P450 or glucuronosyltransferase enzymes) Excretion – the removal of the substance or metabolites from the body. In rare cases, some drugs irreversibly accumulate in body tissue. [citation needed] Some textbooks combine the first two phases as the drug is often administered in an active form, which means that there is no liberation phase. Others include a phase that combines distribution, metabolism and excretion into a disposition phase. Other authors include the drug’s toxicological aspect in what is known as ADME-Tox or ADMET. The two phases of metabolism and excretion can be grouped together under the title elimination. The study of these distinct phases involves the use and manipulation of basic concepts in order to understand the process dynamics. For this reason, in order to fully comprehend the kinetics of a drug it is necessary to have detailed knowledge of a number of factors such as: the properties of the substances that act as excipients, the characteristics of the appropriate biological membranes and the way that substances can cross them, or the characteristics of the enzyme reactions that inactivate the drug. Metrics The following are the most commonly measured pharmacokinetic metrics:[5] The units of the dose in the table are expressed in moles (mol) and molar (M). To express the metrics of the table in units of mass, instead of Amount of substance, simply replace ‘mol’ with ‘g’ and ‘M’ with ‘g/L’. Similarly, other units in the table may be expressed in units of an equivalent dimension by scaling. [6] In pharmacokinetics, steady state refers to the situation where the overall intake of a drug is fairly in dynamic equilibrium with its elimination. In practice, it is generally considered that once regular dosing of a drug is started, steady state is reached after 3 to 5 times its half-life. In steady state and in linear pharmacokinetics, AUCτ=AUC∞. [8] Modeling Models have been developed to simplify conceptualization of the many processes that take place in the interaction between an organism and a chemical substance. Pharmacokinetic modelling may be performed either by noncompartmental or compartmental methods. Multi-compartment models provide the best approximations to reality; however, the complexity involved in adding parameters with that modelling approach means that monocompartmental models and above all two compartmental models are the most-frequently used. The model outputs for a drug can be used in industry (for example, in calculating bioequivalence when designing generic drugs) or in the clinical application of pharmacokinetic concepts. Clinical pharmacokinetics provides many performance guidelines for effective and efficient use of drugs for human-health professionals and in veterinary medicine. Models generally take the form of mathematical formulas that have a corresponding graphical representation. The use of these models allows an understanding of the characteristics of a molecule, as well as how a particular drug will behave given information regarding some of its basic characteristics such as its acid dissociation constant (pKa), bioavailability and solubility, absorption capacity and distribution in the organism. A variety of analysis techniques may be used to develop models, such as nonlinear regression or curve stripping. Noncompartmental analysis Noncompartmental methods estimate PK parameters directly from a table of concentration-time measurements. Noncompartmental methods are versatile in that they do not assume any specific model and generally produce accurate results acceptable for bioequivalence studies. Total drug exposure is most often estimated by area under the curve (AUC) methods, with the trapezoidal rule (numerical integration) the most common method. Due to the dependence on the length of x in the trapezoidal rule, the area estimation is highly dependent on the blood/plasma sampling schedule. That is, the closer time points are, the closer the trapezoids reflect the actual shape of the concentration-time curve. The number of time points available in order to perform a successful NCA analysis should be enough to cover the absorption, distribution and elimination phase to accurately characterize the drug. Beyond AUC exposure measures, parameters such as Cmax (maximum concentration), Tmax (time to maximum concentration), CL and Vd can also be reported using NCA methods.