



# Evaluation of prompt engineering strategies for pharmacokinetic data analysis with the ChatGPT large language model

Euibeom Shin<sup>1</sup> · Murali Ramanathan<sup>1</sup>

Received: 18 September 2023 / Accepted: 11 October 2023 / Published online: 11 November 2023  
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

To systematically assess the ChatGPT large language model on diverse tasks relevant to pharmacokinetic data analysis. ChatGPT was evaluated with prototypical tasks related to report writing, code generation, non-compartmental analysis, and pharmacokinetic word problems. The writing task consisted of writing an introduction for this paper from a draft title. The coding tasks consisted of generating R code for semi-logarithmic graphing of concentration–time profiles and calculating area under the curve and area under the moment curve from time zero to infinity. Pharmacokinetics word problems on single intravenous, extravascular bolus, and multiple dosing were taken from a pharmacokinetics textbook. Chain-of-thought and problem separation were assessed as prompt engineering strategies when errors occurred. ChatGPT showed satisfactory performance on the report writing, code generation tasks and provided accurate information on the principles and methods underlying pharmacokinetic data analysis. However, ChatGPT had high error rates in numerical calculations involving exponential functions. The outputs generated by ChatGPT were not reproducible: the precise content of the output was variable albeit not necessarily erroneous for different instances of the same prompt. Incorporation of prompt engineering strategies reduced but did not eliminate errors in numerical calculations. ChatGPT has the potential to become a powerful productivity tool for writing, knowledge encapsulation, and coding tasks in pharmacokinetic data analysis. The poor accuracy of ChatGPT in numerical calculations require resolution before it can be reliably used for PK and pharmacometrics data analysis.

**Keywords** ChatGPT · Pharmacokinetics · Prompt engineering · Drug development · PK/PD · Graphing · Bioavailability

## Introduction

Large language models (LLM) such as ChatGPT from OpenAI [1], Bard from Google [2], and others are an emerging artificial intelligence technology that has engendered great public interest. LLM are deep learning neural networks trained on a large body of text and other information that are capable of two-way interactions with users in a manner that approximates the complexity and nuance of human conversations.

Pharmacokinetic (PK) and pharmacodynamic (PK-PD) modeling is well established as an effective tool in the pre-

clinical setting for understanding time courses of drug concentrations and effects, for inter-species scaling, and dose determination. Population modeling with non-linear mixed effects and Bayesian methods extends PK-PD modeling. It is particularly useful for interpreting sparse clinical data and for clinical trial simulations to define the scope of drug concentration and effect variability [3, 4]. Every new innovator drug application submitted to the United States Food and Drug Administration contains PK-PD and population modeling data.

PK-PD analyses requires individuals with specialized multi-disciplinary training and utilizes complex software tools with steep learning curves [5, 6]. The model development processes require high levels of human intervention, and the interpretation of results requires expertise and experience. There is a shortage of qualified manpower for the PK-PD analyst workforce needs of industry, regulatory agencies, and academia [5, 6].

---

✉ Murali Ramanathan  
murali@buffalo.edu

<sup>1</sup> Department of Pharmaceutical Sciences, University at Buffalo, The State University of New York, 355 Pharmacy, Buffalo, NY 14214-8033, USA

We reasoned that powerful capabilities of LLM might be potentially harnessed in PK-PD data analyses settings in several ways, e.g., to facilitate self-learning of PK analysis concepts, for identifying the range of available software tools, to provide templates for coding, to conduct uncomplicated PK analyses, and for report preparation. While LLM are reasonably facile at interpreting lay language inputs with a conversational or informal tone, their effectiveness and utility can be further enhanced using structured prompts and exemplars. This has led to research into “prompt engineering” strategies that can guide the underlying LLM into yielding results that more effective for specific classes of tasks. Prompt engineering strategies such as chain-of-thought prompting, which provides the LLM a limited number of examples containing an input, the chain of thought, and the correct output, improves the performance of LLM for solving diverse arithmetic word problems and symbolic reasoning problems [7].

The main goals of this study are to obtain proof-of-concept evidence for utilizing LLM for PK data analysis and to serve as a primer that might motivate pharmaceutical scientists to further explore LLM for more advanced modeling and simulation tasks.

## Methods

### ChatGPT methods

The default version of ChatGPT 4.0 [2] was run at the chat.openai.com on a MacBook Air computer running macOS Ventura 13.5.1. Screenshots from individual ChatGPT runs were saved.

All ChatGPT experiments were replicated  $\geq 3$  times. For experiments involving numerical calculations, the accuracy rate was computed as the number of correct answers based on 10 replicate experiments. Numerical answers from ChatGPT were considered correct if they were within  $\pm 0.5\%$  of the true value; this corresponds to the maximum percent error with an answer expressed to three significant figures. ChatGPT history was cleared between experiments.

Accuracy rates obtained with and without prompt engineering directives were compared. Specific prompt engineering directives were introduced for queries that elicited high error rates. These included the inclusion of phrases such as “Reason it out before action”, “You are a good calculator”, “Provide the most detailed responses”, breaking down multi-part problems into separate questions (i.e., asking one question at a time), and breaking down multi-part problems into separate questions combined with the directive “provide steps”.

### Case study 1. Scientific report generation

In the first experiment, ChatGPT was tasked to generate an outline for a four-paragraph introduction section of a research paper with a title “Evaluation of Prompt Engineering Strategies for PKPD Analysis with ChatGPT”, a draft title that was among those considered for this paper. In the second experiment, ChatGPT was prompted to write a four-paragraph introduction section with references for the research paper with appropriate references.

### Case study 2. Software code generation

Both experiments in this Case Study used concentration–time data shown in Fig. 2A from Patient ID-1 in the vignette in the Ubiquity software package [8].

We requested code from ChatGPT for creating a semi-logarithmic graph and the methods for computing the area under the curve (AUC) and area under the moment curve (AUMC) from zero to infinity. No additional prompt engineering was done, and chat history was cleared between experiments.

AUC and AUMC calculations used the prompt “Provide R code to calculate the AUC and AUMC from time 0 to infinity for a non-compartment analysis for the following concentration(ng/ml) vs. time(hour) profile.

```
[TIME_HR]-(1, 4, 8, 24, 72, 168, 336, 504, 671.9999)
[C_ng_ml]-(9953.8813, 9704.5133, 9383.7171,
8223.5475, 5685.306, 3118.7764, 1673.885, 1215.2236,
964.6353)”.
```

The code generated by ChatGPT was run using the R statistical program [9] in the RStudio environment.

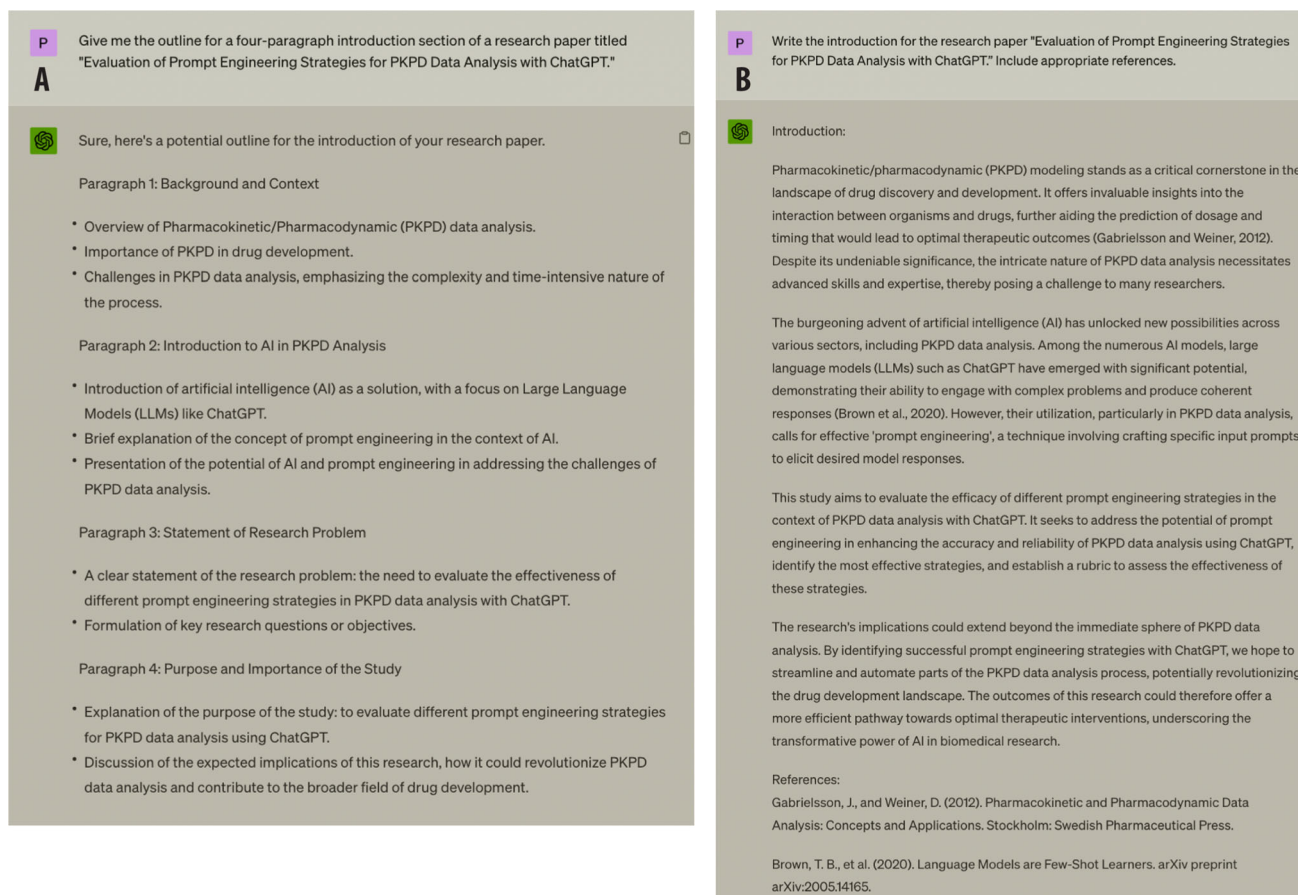
### Case study 3. Pharmacokinetics calculations word problems

The word problems were selected from the textbook *Clinical Pharmacokinetics: Concepts and Applications*, 3rd Edition, by Rowland and Tozer [10]. The answers from ChatGPT were compared to the answer key in the textbook. We assessed the capabilities of ChatGPT at solving word problems involving PK calculations related to intravenous (Chapter 3, Problem 7) and extravascular bolus Chapter 6, Problem 8) dosing, and multiple oral dosing.

## Results

### Case study 1. Scientific report generation

As an LLM, ChatGPT is viewed as a potentially powerful tool to facilitate text generation and writing. In this Case



**Fig. 1** Results from Case Study 1 evaluating the text generation capabilities of ChatGPT. **A** summarizes the ChatGPT prompt and output when tasked to provide an outline for the Introduction section for a research manuscript with a draft version of the title for this

Study, ChatGPT was prompted to provide an outline for the Introduction section of a research paper based on a title and the Introduction section with references. Figure 1 shows screenshots of the ChatGPT output from these two experiments, which were conducted no prompt engineering strategies other than clearly state the task.

The outline produced by ChatGPT is shown in Fig. 1A. The theme and purpose for each paragraph was clearly stated for each of the four paragraphs in the ChatGPT-generated outline. A bulleted list with the main points to be made in each paragraph followed. These results with the outline were considered promising given that limited information, i.e., title only with the PKPD abbreviation, was provided to the LLM.

The Introduction section is shown in Fig. 1B. The generated Introduction section had four paragraphs and hewed to the outline in Fig. 1A. The ChatGPT-generated Introduction section was viewed generally concordant with accepted writing standards in pharmaceutical science

paper. **B** summarizes the corresponding ChatGPT output when it was tasked to write a complete Introduction section with references for a research manuscript with the same title as this paper. The white-on-black screenshots from ChatGPT were recolored to improve contrast

journals and contained two in-text citations and a reference section.

We passed the ChatGPT-generated Introduction section to the Grammarly typing assistance program, which contains a plagiarism check feature. The ChatGPT output was reported as plagiarism free.

The main weaknesses of the ChatGPT-generated Introduction section were its short length and the limited number of references. The short length can be attributed to the prompt statement, which specifically requested only four paragraphs. The Introduction also lacked specificity regarding the PK-PD problems that might benefit from the use of ChatGPT and detail regarding prompt engineering strategies that have been proposed. Because ChatGPT has been reported to make errors in writing, a phenomenon that has been termed “artificial hallucination” [11–13], we checked the references for accuracy. Both references were generally correct. However, the year indicated for the first reference—the Gabrielsson and Weiner text [14, 15]—did

**P** Provide R code for semi-log plot of concentration over time for scientific report.  
[TIME\_HR](1,4,8,24,72,168,336,504,671.9999)  
[C\_ng\_ml](9953.8813, 9704.5133, 9383.7171, 8223.5475, 5685.306, 3118.7764, 1673.885, 1215.2236, 964.6353)

**A**

You can use the 'ggplot2' package in R to generate the semi-log plot. Here is a sample code snippet:

```

r
# Import the required library
library(ggplot2)

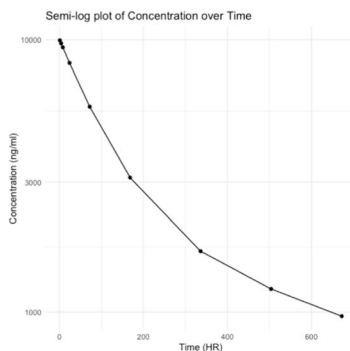
# Define the vectors
TIME_HR <- c(1, 4, 8, 24, 72, 168, 336, 504, 671.9999)
C_ng_ml <- c(9953.8813, 9704.5133, 9383.7171, 8223.5475, 5685.306, 3118.7764, 1673.885, 1215.2236, 964.6353)

# Create a data frame
df <- data.frame(TIME_HR, C_ng_ml)

# Create the semi-log plot
ggplot(data = df, aes(x = TIME_HR, y = C_ng_ml)) +
  geom_point() +
  geom_line() +
  scale_y_log10() +
  labs(x = "Time (HR)", y = "Concentration (ng/ml)",
       title = "Semi-log plot of Concentration over Time") +
  theme_minimal()

```

This script will create a scatterplot with a line connecting the points, where the x-axis represents the time in hours, and the y-axis represents the concentration in ng/ml. The y-axis is in log scale (base 10), making the plot a semi-log plot. The title, x, and y labels are customized, and 'theme\_minimal()' is used to create a clean and minimalistic theme suitable for scientific reports.

**B**

**Fig. 2** Results from Case Study 2 evaluating the code generation capabilities of ChatGPT and to obtain information regarding area under the curve (AUC) and area under the moment curve (AUMC) calculations. **A** summarizes the ChatGPT prompt and output when tasked to provide R code for a semi-logarithmic graph for a

concentration–time data set. **B** shows the graph generated when the code in **A** was executed. **C** summarizes the ChatGPT output when it was tasked to provide an outline of the methods used to compute AUC and AUMC from time zero to infinity. The white-on-black screenshots from ChatGPT were recolored to improve contrast

not match the year of publication of either the 4th or 5th editions of the book. Nonetheless, ChatGPT's proficiency in generating a plausible Introduction for a research manuscript with limited title information and no prompt engineering suggests its potential value as a productivity tool in the writing process.

**P** How to Calculate AUC and AUMC from 0 to infinity.

**C**

The terms AUC (Area Under the Curve) and AUMC (Area Under the Moment Curve) are often used in pharmacokinetics to describe the total exposure to a drug over time and to describe the mean residence time in the body, respectively.

The AUC from 0 to infinity ( $AUC_{0-\infty}$ ) can be estimated by dividing the curve into discrete segments and using numerical methods such as the trapezoidal rule to calculate the area under each segment. If the curve follows exponential decay, you can add the last extrapolated part using the last observed concentration divided by the elimination rate constant.

Here's a step-by-step guide to calculate AUC from 0 to infinity:

- Collect Concentration-Time Data:** Obtain the concentrations of the drug in the blood or other fluid at various times after administration.
- Calculate AUC from 0 to the Last Measured Time Point (t):** Divide the curve into segments and apply a numerical integration method like the trapezoidal rule:
 
$$AUC_{0-t} = \frac{1}{2} \sum_{i=1}^{n-1} (C_i + C_{i+1})(t_{i+1} - t_i)$$
 where  $n$  is the number of time points,  $C_i$  is the concentration at time  $t_i$ , and  $C_{i+1}$  is the concentration at time  $t_{i+1}$ .
- Calculate the Exponential Decay:** Fit an exponential decay function to the last points to determine the elimination rate constant ( $k$ ).
- Calculate the Area from the Last Measured Time Point to Infinity:** Use the last observed concentration and the elimination rate constant:
 
$$AUC_{t-\infty} = \frac{C_{last}}{k}$$
- Add Both Parts Together:**

$$AUC_{0-\infty} = AUC_{0-t} + AUC_{t-\infty}$$

To calculate the AUMC from 0 to infinity, you will follow a similar process, but instead of using the concentrations  $C$ , you will multiply the concentrations by time at each time point:

- Calculate AUMC from 0 to the Last Measured Time Point (t):**

$$AUMC_{0-t} = \frac{1}{2} \sum_{i=1}^{n-1} (C_i \cdot t_i + C_{i+1} \cdot t_{i+1})(t_{i+1} - t_i)$$
- Calculate the Area from the Last Measured Time Point to Infinity:**

$$AUMC_{t-\infty} = \frac{C_{last} \cdot t_{last}}{k} + \frac{C_{last}}{k^2}$$
- Add Both Parts Together:**

$$AUMC_{0-\infty} = AUMC_{0-t} + AUMC_{t-\infty}$$

Note that these calculations are based on assuming certain characteristics about the curve and the way the drug behaves in the body. The details may vary depending on the specific context, so it may be advisable to consult with a pharmacokinetics expert or refer to specialized software designed to handle these calculations.

concentration–time data set. **B** shows the graph generated when the code in **A** was executed. **C** summarizes the ChatGPT output when it was tasked to provide an outline of the methods used to compute AUC and AUMC from time zero to infinity. The white-on-black screenshots from ChatGPT were recolored to improve contrast

## Case study 2. Software code generation

ChatGPT can generate code for programming tasks in several languages including Python and R, which are increasingly used in PK-PD data analysis.

Figure 2A summarizes the prompt and data entered for ChatGPT and the results for the semi-log graphing.



**Table 1** Problems [10] used in Case Study 3 and the accuracy of ChatGPT results

Problem Statement	Accuracy
<b>Chapter 3, Problem 7.</b> When 100 mg of a drug was given as an i.v. bolus, the following plasma concentration-time relationship ( $C$ in mg/L and $t$ in hours) was observed, $C = 7.14e^{-0.051t}$	
a. Volume of distribution.	10/10
b. Elimination half-life.	10/10
c. Total $AUC$ .	10/10
d. Total clearance.	9/10
e. Plasma concentration 70 min after a 250-mg i.v. bolus dose.	6/10

**Chapter 6, Problem 8.** The information in Table 6-6 on cidofovir (Vistide), an antiviral agent, was obtained by Wachsmann et al. (Wachsmann M, Petty PG, Cundy KC, et al. Pharmacokinetics, safety and bioavailability of HPMPC [Cidofovir] in human immunodeficiency virus-infected subjects. Antiviral Res 1996;29:153-161). Use mean values to answer the following questions.

**TABLE 6-6** Area, Half-life, and Amount Excreted Unchanged of Cidofovir after Intravenous, Subcutaneous, and Oral Administration

Dose	1mg/kg		3 mg/kg		10 mg/kg	
Route of administration	i.v.		i.v.	s.c.	i.v.	p.o.
$AUC(0 - \infty)$ ( $\mu\text{g}\cdot\text{hr}/\text{mL}$ )	$7.7 \pm 2.6$		$17.9 \pm 2.3$	$17.8 \pm 1.5$	$64.2 \pm 8.8$	$< 3.5$
Half-life observed (hr)	$1.7 \pm 0.5$		$2.5 \pm 0.9$	$2.3 \pm 0.7$	$2.9 \pm 0.7$	*
Amount excreted unchanged, $Ae(\infty)$ (mg/kg)	$0.92 \pm 0.20$		$2.38 \pm 0.36$	$2.47 \pm 0.25$	$10.5 \pm 1.6$	$0.24 \pm 0.03$

p.o., oral

\* Concentration too low to obtain a proper estimate.

a. Calculate the bioavailability (extent of systemic absorption) of cidofovir when given by the subcutaneous (s.c.) route using: 1.) The plasma data. 2.) The urine data.	10/10 10/10
b. Calculate the bioavailability of cidofovir when given orally using: 1.) The plasma data. 2.) The urine data.	8/10 8/10
c. Estimate the fraction excreted unchanged after all three i.v. doses. 1.) 1 mg/kg 2.) 3 mg/kg 3.) 10 mg/kg	10/10 10/10 10/10

**Chapter 11, Problem 5.** Table 11-8 lists a typical plasma concentration-time profile obtained following an oral 500-mg dose of a drug. The  $AUC$  is 80.6 mg-hr/L, and the terminal half-life is 5 hr.

**TABLE 11-8** Plasma Concentration-Time Profile of a Drug After a Single 500-mg Oral Dose

Time (hr)	0	1	2	4	8	12	24	36	48
Drug Concentration (mg/L)	0	2.3	4.7	5.2	4	2.8	0.6	0.14	0.03

a. What oral dosing rate of drug is needed to maintain an average plateau concentration of 10 mg/L?	8/10
b. The decision has been made to give the drug once every 12 hr. What is: 1.) The unit dose strength of product needed? 2.) The plateau trough concentration expected.	8/10 0/10

ChatGPT recommended the *ggplot2* package [16] and generated R code for creating the graph. The R code ran without errors and generated a line graph (Fig. 2B) titled “*Semi-log plot of Concentration over Time,*” with the *x*-axis and *y*-axis labeled *Time (HR)* and *Concentration(ng/ml)*. The graph esthetics were simple as the minimal theme was recommended in the code. We did not seek to obtain the best fit regression line to an exponential equation for the data.

AUC and AUMC calculations are invariably the first steps in non-compartmental analysis (NCA) of PK data. Some PK domain expertise is required for calculating the AUC and AUMC from time 0 to infinity since parametric interpolation is used to obtain the area of the region between the last observation time point and infinity.

As a first step in our experiment, we investigated whether ChatGPT could provide a detailed description of the methodology needed for calculating AUC and AUMC. The Results in Fig. 2C demonstrate that ChatGPT description is appropriate and complete—the integration and the extrapolation procedures are correctly described using equations where necessary.

Several issues were encountered when conducting AUC and AUMC analyses with concentration–time profile data from Patient ID-1. The ChatGPT-generated R code included useful comments on the procedural steps and ran without syntax errors, but the output was highly variable between occasions even with the same prompt. Despite these variability issues, ChatGPT used appropriate methodology, e.g., integration with the trapezoidal rule, extrapolation from the last observed time point to infinity in every case, which indicates that ChatGPT had generally accessed and identified the correct sources of information for NCA analyses. The accuracy of AUC calculations was 5/10 whereas accuracy of AUMC calculations was 1/10. The elimination rate constant was wrongly calculated in 5/10 experiments, and this was propagated to errors in AUC and AUMC values. In 3 of 10 experiments, ChatGPT assumed that the last two observations were on the terminal phase, in 5 of 10 experiments ChatGPT assumed the last three observations were on the terminal phase, and in one experiment, ChatGPT fit an exponential function to all the data. In every experiment, the remainder of the R code for AUC calculations was correct in terms of trapezoidal rule and extrapolation from the last observed time point to infinity. The corresponding R code for AUMC calculations was correct in 3/10 experiments with most of the errors occurring in the term that extrapolated AUMC from the last time point to infinity.

The results indicate that ChatGPT can generate R code for basic PK analysis tasks such as creating graphs. In calculating metrics such as AUC and AUMC, ChatGPT is

capable of correctly using the trapezoidal rule but is susceptible to range of errors.

### Case study 3. Pharmacokinetics calculations word problems

#### Case study 3A

ChatGPT was posed a five-part single intravenous (IV) bolus dosing PK problem requiring calculation of (a) volume of distribution, (b) elimination half-life, (c) total AUC, (d) total clearance, (e) plasma concentration at 70 min after dose administration.

An 88% accuracy rate was achieved was achieved without prompt engineering (Table 1); 83% of the errors occurred in part (e), which required exponential arithmetic and had an error rate of 50%. Interestingly, ChatGPT set up the exponential equation required for calculation correctly even for incorrect answers. The time in minutes was correctly converted to hours to match the units of the elimination rate constant of the exponential equation.

Prompt engineering efforts were directed exclusively at part (e). The directive “*Reason it out before action*” did not improve accuracy whereas, “*You are an accurate calculator*”, increased accuracy by 6%. When the question in part (e) was posed separately, ChatGPT accuracy increased from 50 to 70%. When further prompted to detail each calculation step, ChatGPT achieved 100% accuracy. However, when the entire question set was presented with the directive to detail each calculation step, accuracy decreased by 6%. This suggests that ChatGPT accuracy improves if calculation problems are compartmentalized.

#### Case study 3B

ChatGPT was posed a seven-part PK problem requiring bioavailability calculations after a single subcutaneous or oral dose. Data on AUC, observed half-life and fraction excreted unchanged in urine following intravenous, subcutaneous, and oral dosing were provided (see Table 1).

Without any prompt engineering, the accuracy of 94.3% (4 errors in 70 experiments) was achieved (Table 1). With the prompt engineering directive “*Reason it out before action*”, the accuracy improved to 98.6% (1 error in 70 experiments). When each question in Case 3B was posed separately or with the directive “*You are an accurate calculator*”, there were no errors (100% accuracy).

#### Case study 3C

This was a 3-part PK problem on multiple oral dosing given the concentration–time profile for a single oral dose.

ChatGPT yielded an accuracy of 80% (2 errors in each of 10 experiments) on parts a and b-1 without any prompt engineering. The accuracy increased to 100% with the directive “*Provide each step of the calculation*”.

ChatGPT did not correctly solve part b-2, which required application of the superposition principle to correctly calculate the multiple dosing concentration profiles from the single oral dosing profile, in any of the 10 experiments. ChatGPT used the equations for a model with linear absorption and one-compartment elimination and made errors in selecting the equation for steady-state trough concentrations. The directive used in parts a and b-1 did not yield accuracy improvements. However, the directive “*Use the superposition principle. The multiple oral dosing profile is the linear superposition of the single oral dose profile at each dose and adjust concentrations based on the ratio of the multiple dose profile to the single dose profile. Assume a steady state approximated happens after 4 doses*” yielded an accuracy of 2/10. With an exemplar based on a correct response, the accuracy only improved to 4/10.

## Discussion

We investigated the strengths and weaknesses of the ChatGPT large language model in PK/PD data analysis and pharmacometrics. In our experimental design, we intentionally included diverse problems representative of the breadth of conceptual learning, quantitative analyses, visualization, coding, and report writing tasks required in the domain. We found that the scientific information and analysis processes recommended by ChatGPT to be appropriate and accurate. There were challenges related to lack of determinism in the algorithm and arithmetic errors in the numerical calculations.

Calculation word problems frequently require arithmetic, commonsense, and symbolic reasoning. Evidence suggests that LLM “struggle” with calculation word problems, which is somewhat unexpected given the satisfactory performance on tasks such as writing and code generation [17, 18]. Interestingly, experiments have shown that the accuracy of ChatGPT in calculations involving exponentiation involving decimals, logarithmic and trigonometric function are only  $\sim 50\%$  [17, 18]. The lack of numerical accuracy with exponentiation and logarithmic operations represents a significant limitation for PK and pharmacometric analyses. We encountered these issues in solving the AUMC and textbook PK calculation problems where we obtained an accuracy of  $\sim 50\%$  for the calculations involving exponentials. Interestingly, in nearly every experiment with the wrong answers, ChatGPT had the correct mathematical expression but the wrong

numerical calculations. We expect that these issues might fully resolve once LLM incorporate a calculator in the algorithm.

The outputs from ChatGPT can vary even when the same prompt is used, which makes it difficult to replicate any particular outcome in exact detail. In stochastic modeling, the lack of determinism is usually addressed by seeding the random number generator. It is not clear that an analogous strategy for resolving this problem is available in this setting because LLM and other generative AI algorithms produce random variates from a complex high dimensional joint distribution [19, 20]. This limitation is problematic in the regulatory setting where reproducibility of results is critical.

Prompt engineering is an approach to transfer knowledge from the user to the LLM [7] and has been shown to improve the performance on arithmetic word problems [21]: e.g., prompting ChatGPT to show its work [22], or adding the chain-of-thought phrase “Let’s think step-by-step” improved the accuracy of GPT-3 on arithmetic problems [23]. We attempted chain-of-thought and problem separation prompt engineering techniques in the cases that we encountered high error rates. While ChatGPT accuracy improved with chain-of-thought prompting on arithmetic problems, we were not able to eliminate numerical calculations errors. This is concordant with the findings of Chen et al. [23] who did not find performance improvement for ChatGPT on arithmetic reasoning problems but noted utility in other problem areas. This was attributed to incorporation of chain-of-thought features into ChatGPT [23].

Given that ChatGPT is an emerging AI tool, there have not been many papers that have specifically investigated its utility in PK analyses and pharmacometrics. One exception is the research by Cloesmeijer et al. [24] who investigated usefulness of ChatGPT for code generation; the R code generated was satisfactory but the NONMEM code contained errors. We investigated R code generation in Case Study 2 but did not investigate NONMEM code here. We also did not conduct calculations involving population PK modeling given the high error rates in the AUC and AUMC calculations. While we found that the R code generated ran without issues in RStudio, thorough reviewing and debugging for conceptual errors was required. It might be useful to evaluate ChatGPT code with a test bank of questions with known answers.

Our results suggest that ChatGPT could be a useful productivity aid for writing, knowledge encapsulation, and programming tasks. However, the high rate of errors in arithmetic calculations could limit its utility for more complex tasks and diverse data analysis scenarios in PK data analysis and pharmacometrics. If the limitations are overcome, ChatGPT could become a valuable tool for

automating all aspects of PK and pharmacometric data analysis. Prompt protocol research and rigorous performance evaluation, and validation studies on benchmark data sets will be required to build the confidence and certainty expected in the pharmaceutical regulatory environment.

**Author contributions** ES—Data analysis, manuscript preparation. MR—Study concept and design, data analysis, manuscript preparation.

**Funding** This is unfunded research. Support from Grant MS190096 from the Department of Defense Multiple Sclerosis Research Program for the Office of the Congressionally Directed Medical Research Programs (CDMRP) to the Ramanathan laboratory is gratefully acknowledged.

## Declarations

**Conflict of interest** Euibeom Shin has no conflicts. Dr. Murali Ramanathan received research funding from the National Multiple Sclerosis Society, Department of Defense, National Science Foundation, and National Institute of Neurological Diseases and Stroke. He receives royalty from a self-published textbook.

## References

- OpenAI (2023) ChatGPT (June 26 version) Large language model
- Google AI (2023) Bard Large language model
- Kimko HC, Duffull SB (2003) Simulation for designing clinical trials: a pharmacokinetic-pharmacodynamic modeling perspective: the pharmaceutical sciences, vol 127. Marcel Dekker, New York
- Kimko HC, Peck CC, American Association of Pharmaceutical Scientists (2011) Clinical trial simulations: applications and trends. AAPS advances in the pharmaceutical sciences series, vol 1. AAPS Press, Springer, New York
- Bonate PL, Barrett JS, Ait-Oudhia S, Brundage R, Corrigan B, Duffull S, Gastonguay M, Karlsson MO, Kijima S, Krause A, Lovern M, Neely M, Ouellet D, Plan EL, Rao GG, Standing J, Wilkins J, Zhu H (2023) Training the next generation of pharmacometric modelers: a multisector perspective. *J Pharmacokinet Pharmacodyn*. <https://doi.org/10.1007/s10928-023-09878-4>
- Michelet R, Aulin LBS, Borghardt JM, Costa TD, Denti P, Ibarra M, Ma G, Meibohm B, Pillai GC, Schmidt S, Hennig S, Kloft C (2023) Barriers to global pharmacometrics: educational challenges and opportunities across the globe. *CPT Pharmacometrics Syst Pharmacol* 12(6):743–747. <https://doi.org/10.1002/psp4.12940>
- White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, Elnashar A, Spencer-Smith J, Schmidt DC (2023) A prompt pattern catalog to enhance prompt engineering with ChatGPT. [arXiv:2302.11382](https://arxiv.org/abs/2302.11382)
- Harrold JM, Abraham AK (2014) Ubiquity: a framework for physiological/mechanism-based pharmacokinetic/pharmacodynamic model development and deployment. *J Pharmacokinet Pharmacodyn* 41(2):141–151. <https://doi.org/10.1007/s10928-014-9352-6>
- R Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
- Rowland M, Tozer TN (1995) Clinical pharmacokinetics: concepts and applications, 3rd edn. Williams & Wilkins, Baltimore
- Alkaiissi H, McFarlane SI (2023) Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus J Med Sci* 15(2):e35179. <https://doi.org/10.7759/cureus.35179>
- Athaluri SA, Manthena SV, Kesapragada V, Yarlagadda V, Dave T, Duddumpudi RTS (2023) Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus* 15(4):e37432. <https://doi.org/10.7759/cureus.37432>
- Beutel G, Geerits E, Kielstein JT (2023) Artificial hallucination: GPT on LSD? *Crit Care* 27(1):148. <https://doi.org/10.1186/s13054-023-04425-6>
- Gabrielsson J, Weiner D (2007) Pharmacokinetic–pharmacodynamic data analysis: concepts and applications, 4th edn. Swedish Pharmaceutical Press, Stockholm
- Gabrielsson J, Weiner D (2016) Pharmacokinetic–pharmacodynamic data analysis: concepts and applications, 5th edn. Swedish Pharmaceutical Press, Stockholm
- Wickham H (2009) ggplot2: elegant graphics for data analysis. *Use R*:1–212. <https://doi.org/10.1007/978-0-387-98141-3>
- Frieder S, Pinchetti L, Chevalier A, Griffiths R-R, Salvatori T, Lukaszewicz T, Petersen PC, Berner J (2023) Mathematical capabilities of ChatGPT. [arXiv:2301.13867v13862](https://arxiv.org/abs/2301.13867v13862)
- Yuan Z, Yuan H, Tan C, Wang W, Huang S (2023) How well do large language models perform in arithmetic tasks? [arXiv:2304.02015](https://arxiv.org/abs/2304.02015)
- Nair R, Mohan DD, Frank S, Setlur S, Govindaraju V, Ramanathan M (2023) Generative adversarial networks for modelling clinical biomarker profiles with race/ethnicity. *Br J Clin Pharmacol* 89(5):1588–1600. <https://doi.org/10.1111/bcp.15623>
- Nair R, Mohan DD, Setlur S, Govindaraju V, Ramanathan M (2023) Generative models for age, race/ethnicity, and disease state dependence of physiological determinants of drug dosing. *J Pharmacokinet Pharmacodyn* 50(2):111–122. <https://doi.org/10.1007/s10928-022-09838-4>
- Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi EH, Le QV, Zhou D Chain-of-thought prompting elicits reasoning in large language models. In: 36th conference on neural information processing systems (NeurIPS 2022). New Orleans, LA, 2022. vol 1. NeuroIPS Foundation, pp xvi, 538
- Shakarian P, Koyyalamudi A, Ngu N, Mareedu L (2023) An independent evaluation of ChatGPT on mathematical word problems (MWP). [arXiv:2302.13814v13812](https://arxiv.org/abs/2302.13814v13812)
- Chen J, Chen L, Huang H, Zhou T (2023) When do you need chain-of-thought prompting for ChatGPT? [arXiv:2304.03262v03262](https://arxiv.org/abs/2304.03262v03262)
- Cloesmeijer M, Janssen A, Koopman S, Cnossen M, Mathot R (2023) ChatGPT in pharmacometrics? Potential opportunities and limitations

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



## Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval, sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

[onlineservice@springernature.com](mailto:onlineservice@springernature.com)