FAST TRACK

Generative AI and simulation modeling: how should you (not) use large language models like ChatGPT

Ali Akhavan[a] and Mohammad S. Jalali[a,b*]

Abstract

Generative Artificial Intelligence (AI) tools, such as Large Language Models (LLMs) and chatbots like ChatGPT, hold promise for advancing simulation modeling. Despite their growing prominence and associated debates, there remains a gap in comprehending the potential of generative AI in this field and a lack of guidelines for its effective deployment. This article endeavors to bridge these gaps. We discuss the applications of ChatGPT through an example of modeling COVID-19's impact on economic growth in the United States. However, our guidelines are generic and can be applied to a broader range of generative AI tools. Our work presents a systematic approach for integrating generative AI across the simulation research continuum, from problem articulation to insight derivation and documentation, independent of the specific simulation modeling method. We emphasize while these tools offer enhancements in refining ideas and expediting processes, they should complement rather than replace critical thinking inherent to research.

Syst. Dyn. Rev. **40**, e1773 (2024)

Additional Supporting Information may be found online in the supporting information tab for this article.

Introduction

Generative Artificial Intelligence (AI) and Large Langue Models (LLMs), for example, tools like ChatGPT, have received widespread recognition from researchers. While their role in writing and reviewing manuscripts has been extensively discussed and scrutinized (e.g. Ariyaratne *et al.*, 2023; Castellanos-Gomez, 2023), their impact on the overall research process has remained controversial.

Generative AI holds immense potential for research while posing challenges and concerns. On the one hand, researchers argue that generative AI often yields little in terms of actual research outputs. They suggest that tools like ChatGPT produce "seemingly credible but incorrect responses" (Shen *et al.*, 2023) and simply provide "endless entertainment" (Thorp, 2023, p. 313). In the early days of ChatGPT, it generated and referenced scientific studies that did not exist (Thorp, 2023), a similar issue across other platforms too. Meta's Galactica produced inaccurate and racist content (Stokel-Walker and Van Noorden, 2023). In our experience, we observed that Google's Bard fabricated peer-reviewed article titles while claiming they were pulled out of Google Scholar. Consequently, several major journals, including but not limited to *Nature* and *Science*, released guidance on the responsible use of generative AI, and *Nature* (Nature

[a] MGH Institute for Technology Assessment, Harvard Medical School, Boston, Massachusetts, USA
[b] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

* Correspondence to: Mohammad S. Jalali, MGH Institute for Technology Assessment, 125 Nashua St., Boston, MA 02114 USA E-mail: msjalali@mgh.harvard.edu

Accepted by Andreas Größler, Received 21 February 2024; Revised 2 April 2024; Accepted 23 April 2024

System Dynamics Review
System Dynamics Review vol 40, No 3 (July/September 2024): e1773
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/sdr.1773

Editorial, 2023), among several others, prohibited the inclusion of generative chatbots as authors because AI cannot take responsibility for the originality and integrity of the work (Flanagin *et al.*, 2023; Thorp, 2023).

On the other hand, some argue that generative AI can contribute to the research process. They suggest that generative AI offers great potential to improve academic work efficiency (Meyer *et al.*, 2023), and tools like chatbots enable the exploration of "uncharted scientific territories" (Thapa and Adhikari, 2023, p. 2647; Ghaffarzadegan *et al.*, 2024). For example, ChatGPT generates research ideas, literature reviews, and data summaries (Dowling and Lucey, 2023). These tools help researchers revise and review manuscripts and develop or check codes (Stokel-Walker and Van Noorden, 2023). It is also proposed that they may help flatten the disparities among research institutes by providing relatively cheap but powerful resources (Wang *et al.*, 2023).

The controversy surrounding the value of generative AI's applications and their direct (e.g. first-hand research process) and indirect (e.g. providing information resources) impacts on the research process continues. There remains a lack of clarity about the technology's role in carrying out research, especially in simulation modeling.

In this article, we aim to clarify and demonstrate how to use generative AI when conducting simulation modeling research. We also discuss how to harness the power of generative AI in translating verbal and written prompts to outcomes assisting the research process. We use a simplified case of COVID-19 onset and its impact on GDP in the United States. Particularly, we consider the impact of COVID-19 as an exogenous factor affecting a simplified consumption multiplier model (Samuelson, 1939; Low, 1980). Finally, the case provides a realistic context where the applications of generative AI can be tested in a practical way.

Throughout this article, we show and emphasize that generative AI should not replace thinking; instead, it is a useful tool to facilitate the research process, a practical way to review the content generated by researchers, and an enhancement of idea implementation in simulation modeling.

Background and procedures

Large Language Models (LLMs) are deep-learning models trained to understand and generate human language. ChatGPT and Bard, which use generative transformer architecture, are examples of such models that belong to the wider family of LLMs. These models are trained on vast amounts of text data to predict the next word in a sequence, and they can be used to generate coherent and contextually relevant content. While LLMs are a subset of generative AI that focuses on text, generative AI can also include models that generate images, music, videos, or other forms of content.

Recently, different AI platforms have started offering combination capabilities. Platforms such as Google's Bard, Microsoft's Bing Chat, and OpenAI's ChatGPT, among others, are known as conversational AI models that provide a mixture of LLMs and other generative AI tools, such as image processing and advanced data analysis, making them a useful instrument in the research process. Although the growth and development of these tools have been rapid over the recent years and

DOI: 10.1002/sdr

months, GPT-4 (the most recent version) offers the most capabilities at the time of writing this article. Therefore, we focus on the applications of GPT-4 while acknowledging that more evolved technologies may be introduced in the near future.

The advancement and evolution of such AI technologies should not change the fundamentals of scientific inquiries and rigor. Several scholars (e.g. Banks, 2005; Davis *et al.*, 2007; Law *et al.*, 2007; Sterman, 2000) have proposed steps of the modeling process in the simulation research, and the iterative nature of this process has been widely emphasized in the literature (e.g. Harrison *et al.*, 2007; Homer, 1996). We follow the same approach, putting ourselves in the shoes of a researcher conducting simulation research and getting feedback from ChatGPT. We also extend the steps in the modeling process to proper model documentation and maintenance to enhance transparency and replicability of the simulation research, an area that still has significant room for improvement (Jalali *et al.*, 2021; Jalali and Beaulieu, 2023; Monks *et al.*, 2019; Rahmandad and Sterman, 2012).

In the next section, we go through several steps of modeling as we present a case study. We build upon the frameworks on how to conduct simulation modeling research and develop theories from case studies (Davis *et al.*, 2007; Sterman, 2000). We considered general steps that are applicable to most simulation modeling approaches rather than focusing on one specific method and show how ChatGPT can help shape and refine the process. It is important to note that simulation modeling is not a linear, step-by-step process; rather, it often involves looping back and revisiting earlier stages as new insights emerge or challenges are encountered. In other words, the steps are interdependent and iterative. Similarly, during the writing of this article and our deepening understanding of the problem, we found ourselves revisiting some of the earlier steps for improvement.

Additionally, since generative AI tools, like ChatGPT, rely on stochastic models (i.e. heterogeneity in how sentences are formed), they should receive clear instructions (prompts) from users. However, the absence of established guidelines for crafting optimal prompts may lead to the generation of irrelevant outputs. In such cases, we revised prompts to achieve satisfactory responses, known as prompt engineering (e.g. White *et al.*, 2023a). Thus, one contribution of this article is providing well-crafted examples of prompts that illustrate effective communication with generative AI. These examples serve as practical guides for users to understand how to formulate prompts that elicit meaningful responses from tools like ChatGPT.

Below, we first present these prompts and ChatGPT's corresponding responses. ChatGPT's responses are often long; therefore, to keep this article concise, we summarized its responses. We then discuss essential aspects to consider in the creation and enhancement of prompts, for example, iterative process to refine the prompts or ethical considerations.

Using generative AI in simulation modeling — case study

Here, we demonstrate the use of generative AI in simulation research through an interactive dialogic approach. Our prompts are crafted from the perspective of a

DOI: 10.1002/sdr

student or a junior researcher at the outset of their journey in simulation modeling. This method is designed to showcase the utilization of these tools and assist in the educational process of refining the various steps involved in modeling. While we intentionally introduce a few mistakes (we flag such intentional errors with [ ] to avoid confusion), from simple errors to more complex ones, to assess ChatGPT's ability to identify and correct the issues, our approach goes beyond error correction. Through this dialog-based approach, we aim to provide insights for learners, demonstrating how generative AI can be a useful tool for enhancing understanding and fostering analytical skills in simulation research.

Problem definition and research question

Generative AI can help refine research questions and explore problem background. It can understand the breadth and scope of research questions and analyze the semantics and causal relationships between research question variables. Generative AI connected to the internet can search the web and provide preliminary literature or data sources, facilitating research initiation. They can also help with research boundary considerations by suggesting aspects researchers may not have considered. We show this through the example below.

Research question and topic

Prompt

I am starting simulation modeling research and want to cross-check my research question. I have considered this question: How has economic growth impacted COVID-19? You should be critical, find potential flaws in my research question, and suggest improvements. [We intentionally reversed the causality and considered broad terms instead of being specific.]

Response summary

ChatGPT provided feedback on research question clarity and directionality, causal relationships, operational definitions, and temporal dynamics. Regarding the research question's clarity and directionality, it noted that our "research question seems to suggest that economic growth has influenced the virus itself, which might not be the intended meaning. It's more likely you're trying to explore how economic growth has been affected by the pandemic or how economic growth has influenced the spread, containment, or effects of the pandemic. Depending on your intention, you might need to reword for clarity."

Regarding causal relationships, ChatGPT noted to differentiate between causality and correlation. It suggested that "Saying economic growth 'impacted' COVID-19 assumes a causal relationship. However, in reality, many variables interplay, and determining a direct causal relationship might be challenging." In addition, ChatGPT asked us to specify the temporal dynamics of our research. It specifically elaborated that "When you mention economic growth, are you referring to the growth that occurred before the pandemic, during it, or after? This needs to be defined."

Furthermore, it asked us about the operational definition of "economic growth" and "impact." It specifically mentioned: "Are you referring to GDP growth, stock

DOI: 10.1002/sdr

market performance, job creation, or another measure? By 'impact,' do you mean the effect on case numbers, mortality rates, policy decisions, or something else?"

Revised research question

Prompt

These are good points. I see that I made a mistake in the direction of causality and that I was not specific enough. It should have been the impact of COVID-19 shock on economic growth. Regarding operationalization, I am interested in focusing on GDP growth. Regarding the temporal dynamics, I am considering the GDP trend from 2013 to 2023, and I want to focus on the U.S. economy. With those aspects, my revised research question is: How did changes in consumer behavior during the COVID-19 pandemic impact U.S. GDP growth? How does that sound?

Response summary

ChatGPT acknowledged the changes we made in the research question and provided feedback such as "By specifying the years 2013 to 2023, you're including a period before the emergence of COVID-19, which is good. This will allow you to establish a baseline trend before the pandemic and then assess any deviations from this trend during and possibly after the pandemic." Additionally, it provided feedback on counterfactual scenarios, confounding factors, and data sources. For example, it suggested considering what the GDP growth would have been in the absence of the pandemic or how other confounding factors, such as fiscal policy, might have played roles in the U.S. GDP growth. Finally, it recommended accessing data from sources like the U.S. Bureau of Economic Analysis.

Preliminary review of the literature and data sources

Prompt

Can you search for similar peer-reviewed journal articles using simulation research? Can you also introduce a few other reliable sources so that I can get data on GDP growth and relevant economic measures?

Response summary

Unlike earlier versions of chatbots, which fabricated citations that did not exist, recent versions of ChatGPT provide relevant literature with links to published material. In response to our query, ChatGPT provided a relevant journal article (Xiang *et al.*, 2021). Furthermore, it noted databases such as the U.S. Bureau of Economic Analysis and St. Louis Federal Reserve (FRED) that report GDP growth data. ChatGPT also provided a brief overview of the content of the studies. For example, it said that the given article "discusses a model that interlinks the dynamics of labor supply in the economy with the spread of infectious diseases like COVID-19," and about the database it said, "St. Louis Federal Reserve (FRED): Offers historical GDP data from Q1 1950 to Q2 2023, which could be very useful for analyzing the trends you're interested in."

Through further follow-up prompts, we asked for more specific details from the reported reference that we do not report here. What ChatGPT provides is not exhaustive; therefore, such a process can be conducted more systematically to assist with targeted literature search and complement manual searches in databases.

DOI: 10.1002/sdr

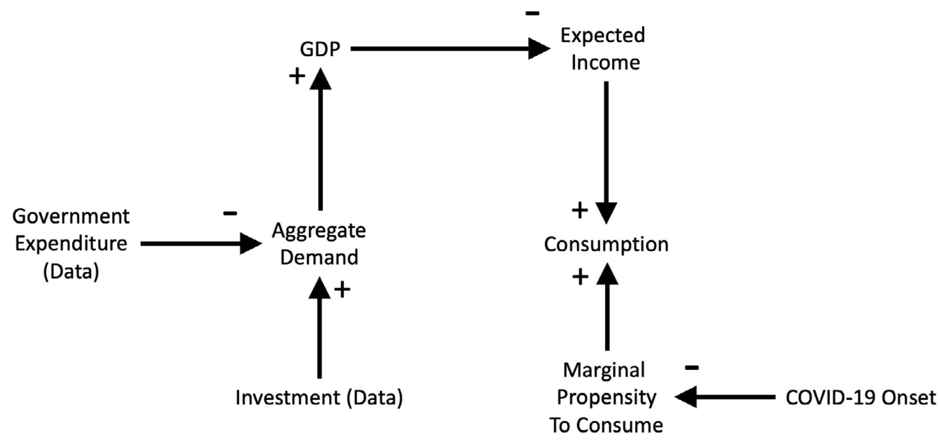


Fig. 1. Initial model of consumption multiplier, provided in a ChatGPT prompt

Model conceptualization

Generative AI can help refine conceptual models. Recent developments enable generative AI to analyze complex visual representations (e.g. images) of conceptual models and identify the relationships among the identified variables. We used an image in our prompt below (Figure 1); however, the content of the image could be simply presented in text too. Generative AI can provide feedback about the choice of variables and their naming, direction of causality, polarity of the relationships, and missing causal links and variables, among others. Generative AI can also check the definition of model variables and suggest new ones.

Conceptual model

Prompt

I want your feedback on my attempt to identify variables and develop a conceptual model for including COVID-19 shock in a simplified macroeconomic model of the consumption multiplier. I have put together the key relationships for my model. Can you check the attached photo and give me your feedback? Be critical. [We intentionally made mistakes in the conceptual model by removing a few variables and causal relationships.]

Response summary

ChatGPT reviewed our variables one by one and identified the relationships, their directionalities, and the links' polarities. It then identified the mistakes we made intentionally while acknowledging the correct aspects. It recognized the incorrect causal polarity between GDP and expected income by indicating that "as GDP grows, we expect the average income or expected income of the population to rise." It also identified the wrong relationship between government expenditure and aggregate demand by specifying: "This is unusual. Generally, an increase in government expenditure, all else being equal, would lead to an increase in aggregate demand."

In addition, ChatGPT offered additional variables, suggesting factors such as exports and imports that are common in macroeconomic models of GDP. It

DOI: 10.1002/sdr

recognized the lack of a link from consumption to aggregate demand, creating a feedback loop: "In reality, an increase in aggregate demand might lead to an increase in GDP, which then affects expected income and thus consumption, creating a cyclical effect." It also reminded us about the complexity and simplification trade-off, such that while it is important to keep the model simple, we should ensure that it still captures the necessary complexity of the real-world problem. Finally, it reminded us about stating our assumptions in the model.

As noted earlier, ChatGPT's responses are often long. We provided its full response for this prompt in Appendix S1 in the online supporting information.

Revised conceptual model

Prompt

I have revised my conceptual model and have attached the updated version [-Figure 2]. *I fixed the polarity errors. I considered the impact of Net Exports on GDP in addition to Aggregate Demand. I also added the causal link between Consumption and Aggregate Demand. For simplicity, I have not included parameters such as initial values and adjustment times, and the conceptual model only captures the basic dynamics. I want to move forward with this version if there are no errors.*

Response summary

ChatGPT understood the changes we made in the model, reviewed the causal relationships one by one, and provided feedback about the model's alignment with already established economic theories. It denoted that "The model is cleanly presented with easy-to-follow links and clear directionality indicators. It should be understandable to those familiar with basic macroeconomic principles." It noted an important observation that the model misses time lags: "While not represented in the simplified model, be mindful of this." It also reminded us that the model does not capture all potential causal relationships, which aligns with our intention of narrowing the boundary for this simple model. On that point, it also noted "While your model is simplified, be aware that in a more dynamic model, changes in GDP can feed back into government expenditure, investment, and net

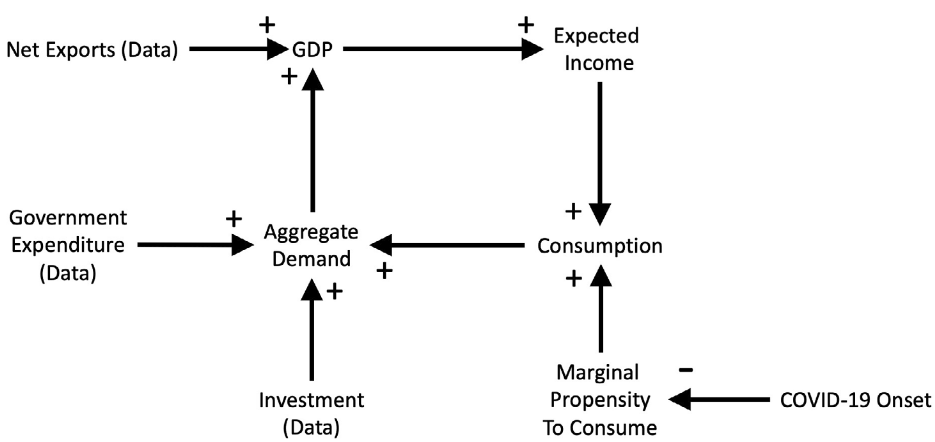


Fig. 2. Revised conceptual model provided in a ChatGPT prompt

DOI: 10.1002/sdr

exports, creating complex interactions." This is a great point; however, we did not consider those additional impacts given that we use exogenous data for the three variables, and as such, their historical effects are already captured. If the model is used for future projections, the inclusion of those feedback loops is essential, but that is out of the scope of this current simple model. Finally, ChatGPT reminded us that all models are simplifications of reality, and we should check to ensure that our model remains reasonable under various economic conditions or policy regimes.

Model formulation and implementation

Generative AI can provide feedback on model equations and parameter values. They can check whether the defined equations are consistent and that there is no logical flaw in formulating the model. They can also inform if the defined parameters are in reasonable ranges. Additionally, generative AI tools are great for translating written mathematical equations and verbal communications into programming languages. These tools provide a great learning opportunity by providing programming guidelines that can facilitate the code-development process and translate code from one language to another. Finally, one can modify codes in real time by communicating with generative AI about their desired changes to the codes.

Choice of the modeling approach

Prompt

I want to develop a simulation model based on my conceptual model and observe the behavior of the variables over time at an aggregate level. Based on what I have learned in my classes and some initial searches, I think a compartmental modeling approach would be a reasonable choice. What do you think?

Response summary

ChatGPT's response emphasized the suitability of using a compartmental modeling approach, typically seen in epidemiology and systems dynamics, for macroeconomic modeling. It outlined several advantages: the model's ability to represent economic variables as "stocks" and "flows," thus capturing the dynamics of aggregate demand, supply, income, and consumption; the capacity to understand the dynamic behavior of economic systems, including delays and feedback loops; its utility in simulating policy interventions and their impacts over time; and the ability to incorporate nonlinearity and complexity, crucial for depicting interactions in an economic system. However, ChatGPT also noted limitations: the potential oversimplification of complex economic interactions; the criticality of accurate parameterization; the necessity for model validation and calibration against real-world data; the challenge of capturing human behavior and expectations; and the need for careful translation of simulation results into real-world policy. ChatGPT concluded that while a compartmental model is a good starting point for observing economic variable behavior over time, especially in response to shocks like COVID-19, a more extensive analysis may require integrating this approach with other modeling approaches.

DOI: 10.1002/sdr

Model formulations and parameter values

Prompt

Thanks! I feel more confident about my decision to choose a compartmental modeling approach. I have defined the equations below and am trying to use ordinary differential equations to model the dynamic relationships among the variables. For this draft, I have considered a hypothetical case and assigned arbitrary numbers to parameters. Be critical and give me feedback on the equations and parameter values. [We intentionally used a wrong equation for "Aggregate Demand" and an unusually high value for "Marginal Propensity to Consume."]

- Expected Income = INTEGRAL (Change in Expected Income, Initial Expected Income).
- Consumption = Marginal Propensity to Consume × Expected Income.
- Change in Expected Income = (GDP−Expected Income) / Expectation Formation Time.
- GDP = INTEGRAL (Change in GDP, Initial GDP).
- Aggregate Demand = Consumption × Investment × Government Expenditure × Net Exports.
- Change in GDP = (Aggregate Demand + Net Exports−GDP) / Production Adjustment Time.
- Expectation Formation Time = 2.
- Initial Expected Income = 250.
- Initial GDP = 250.
- Government Expenditure = 80.
- Investment = 10.
- Net Export = 10.
- Marginal Propensity to Consume = IF 2020 ≤ Time ≤ 2020.25 THEN Reference Marginal Propensity to Consume × (1−Alpha) ELSE Reference Marginal Propensity to Consume.
- Production Adjustment Time = 1.
- Reference Marginal Propensity to Consume = 80.
- Alpha = 0.

Response summary

Overall, ChatGPT acknowledged that using ordinary differential equations (ODEs) is a good start and reminded us to ensure that our simulation environment or software can correctly interpret our integral formulations. It reviewed and commented on all equations and identified the mistake we made in the definition of aggregate demand, indicating that "the aggregate demand is the product of its three components. In most macroeconomic models, aggregate demand is the sum of its components. You might want to change this to an additive model." It noted that parameter values seem arbitrary and commented on their intuitions. For instance, it noted that the value of expectation formation time should reflect how quickly or slowly agents update their expectations in response to changes in GDP. It also recognized the other error, suggesting "note that the Marginal Propensity to Consume is typically a fraction (between 0 and 1). In your parameters, you have MPC = 80, which suggests that for every unit of expected income,

DOI: 10.1002/sdr

consumption increases by 80 units. This is not typical; consider revising this to a value between 0 and 1." It also added that overly simplistic models might not capture the nuances of economic relationships, while overly complex models can be difficult to analyze. Finally, it asked us to ensure that all equations are coherent and that the model as a whole makes sense. Each component should logically fit within the broader framework of the model.

Revised equations and variable units

Prompt

I made a mistake in defining Aggregate Demand and setting the parameter value of Marginal Propensity to Consume. I have revised those and considered the additive formulation for Aggregate Demand and 0.80 for MPC. I also added the units of measure for each variable below. The units seem fine to me. Can you check my revised equations and for unit consistency? [We intentionally used wrong units for "Government Expenditure" and "Change in GDP."]

Equations: [for presentation in this article, we only show the modified equations.]

- Aggregate Demand = Consumption + Investment + Government Expenditure + Net Exports.
- Reference Marginal Propensity to Consume = 0.8.

Units:

- Expected Income: $/Year.
- Consumption: $/Year.
- Change in Expected Income: $/(Year × Year).
- GDP: $/Year.
- Aggregate Demand: $/Year.
- Change in GDP: $/Quarter.
- Expectation Formation Time: Year.
- Initial Expected Income: $/Year.
- Initial GDP: $/Year.
- Government Expenditure: $.
- Investment: $/Year.
- Net Exports: $/Year.
- Marginal Propensity to Consume: Dimensionless.
- Production Adjustment Time: Year.
- Alpha: Dimensionless.

Response Summary

ChatGPT reviewed and verified the revised equations. It checked for consistency of the units of measure based on our mathematical formulations. It identified inconsistent units: "The unit of Government Expenditure should be $/Year to be consistent with the Aggregate Demand equation." It also recommended that "The unit of Change in GDP should be adjusted to match its equation." It went

DOI: 10.1002/sdr

through calculations and mentioned "The equation gives units of $/(Year × Year), but you've listed the units as $/Quarter. Either adjust the units in your list or reconsider the equation's formulation."

Developed code

Prompt

I fixed the incorrect units. I then developed the Python code below for my ordinary differential equations [for presentation in this article, we report the code in Appendix S2 in the online supporting information]. *I want the model to start from 2013 to the end of 2022 and the model updates to happen at time intervals of 0.25. I run the model; it works, and I can see the results. Can you check if there are any issues with the code? Anything to improve? Do not provide revised code; instead, tell me how to improve it piece by piece, as I would like to learn how to improve.*

Response summary

ChatGPT reviewed the code and provided comments about documentation, variable naming conventions, equations grouping, model time frame, and units in the plot we created to show the results. It mentioned that our code lacks comments explaining the purpose of functions and variables, making it hard to understand the logic and flow. It also mentioned several variable names are abbreviations that may not be immediately clear to someone reading the code and suggested using full descriptive names. In addition, ChatGPT noted that the equations are implemented directly inside the "model" function without separation and suggested defining separate functions for each equation or group of related equations for better readability and maintenance. Furthermore, it mentioned that we specified the time range as *np.arange* (*2013*, *2023*, *0.25*), which does not include the end of the last quarter of 2022. We should instead extend the range to 2023.25 to model the entire period up to and including the end of 2022, as the quarters run until the start of 2023. It provided useful hints, such as exporting the data for further analysis, and provided the respective Python function. It also noted that the initial conditions should match those of 2013 and emphasized the importance of checking the model against real-world data and conducting sensitivity analysis; we will consider these in later stages below.

Model testing and analysis

Generative AI can help facilitate evaluating and assessing simulation models. They can also provide data analysis and interpretation, suggest model refinements, facilitate model calibration and optimization, and help with sensitivity analysis. While they can simulate and analyze code, they can interpret model results in relation to data (e.g. in Figure 3) and provide suggestions for enhanced model analysis and verification processes.

Comparing with data

Prompt

Below, *I include the updated version of my code based on your feedback* [see Appendix S2 in the online supporting information]. *Now*, *I'm trying to assess the*

DOI: 10.1002/sdr

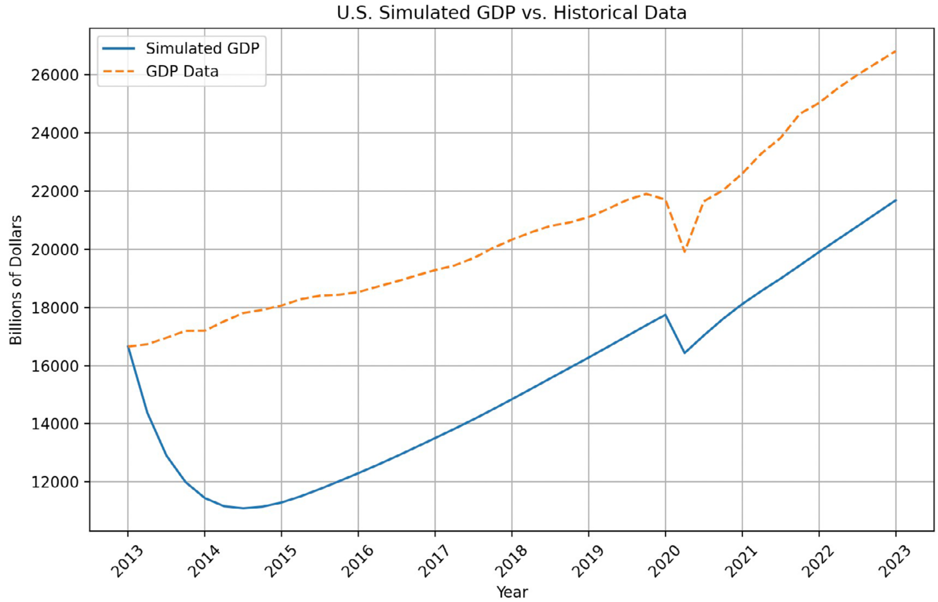


Fig. 3. Using data variables in the model, provided in a ChatGPT prompt

model. With my model variables being based on their historical data, I expect that my model outcome, that is, simulated GDP, be reasonably close to the historical trends of GDP. But when I plot the Simulated GDP against historical data, I don't see a close replication of the historical trend. I am sure about the data quality, though, and I already checked for unit consistency. Give me feedback based on the code and the attached results figure. I aim to learn how to improve the model to replicate the historical trend closely.

Response summary

ChatGPT provided feedback on COVID-19 shock modeling, data interpolation, parameter calibration, model equations, and sensitivity analysis. First, it asked us to ensure the COVID-19 shock is appropriately modeled and the assumptions apply. The model should reflect not only changes in Marginal Propensity to Consume, but also other economic factors affected by the pandemic, such as investment, government expenditure, and net exports. (In our initial model conceptualization in Figure 1, we only mentioned these three variables without "data" in their names; however, upon receiving feedback at this stage, we revised the figure to explicitly include that we utilize historical data. Subsequently, we replicated the following prompts based on this updated understanding.) Second, it asked us to double-check the interpolation methods used for time series data and ensure that the extrapolation beyond the dataset is reasonable and aligns with economic expectations. Third, ChatGPT suggested reviewing parameters like Expectation Formation Time, Production Adjustment Time, and Reference Marginal Propensity to Consume to ensure they are calibrated against historical crises and recovery periods and are consistent with the literature. Fourth, it recommended examining the system of Ordinary Differential Equations to ensure they capture the complex dynamics of an economy during shock periods. It also suggested considering adding additional equations or feedback mechanisms if necessary.

DOI: 10.1002/sdr

Interestingly, it noted that the timeframe of 2020–21 might require more granular data than quarterly due to rapid changes by the pandemic. However, we skipped the suggestion for simplicity of this exercise. Finally, ChatGPT asked to conduct a sensitivity analysis of the key parameters to understand their impact on the model's output and refine them based on historical performance.

Initial model calibration

Prompt

I did not notice that I should have calibrated model parameters; I thought including the exogenous time series for model variables would be sufficient. But good catch, I now understand why the fit was not closely replicating the GDP data trend. I extended the code to estimate all model parameters with their corresponding ranges:

- Expectation Formation Time (0.25, 10).
- Production Adjustment Time (0.125, 2).
- Alpha (0,1).
- Reference MPC (0.5, 1).
- Initial Expected Income (13,048.2, 17,648.2).

I have included my code [presented in Appendix S2 in the online supporting information]. *How would I know if the optimization yielded a good fit?*

Response summary

ChatGPT provided feedback on assessing the effectiveness of an optimization model in achieving a good fit. Key factors include the objective function value, where a lower typically indicates a better fit for methods like least squares; goodness-of-fit metrics such as R-squared, where its higher value (closer to 1) suggests a better fit; the reasonableness and range adherence of optimized parameter values within the model's context; the comparison of model predictions versus actual data, often visualized using tools (like matplotlib in Python) to gauge fit quality; convergence information from the optimization algorithm, including whether a solution was reached and the number of iterations taken; and sensitivity analysis to understand the robustness of the fit by examining how changes in parameter values affect the model's output.

Goodness-of-fit

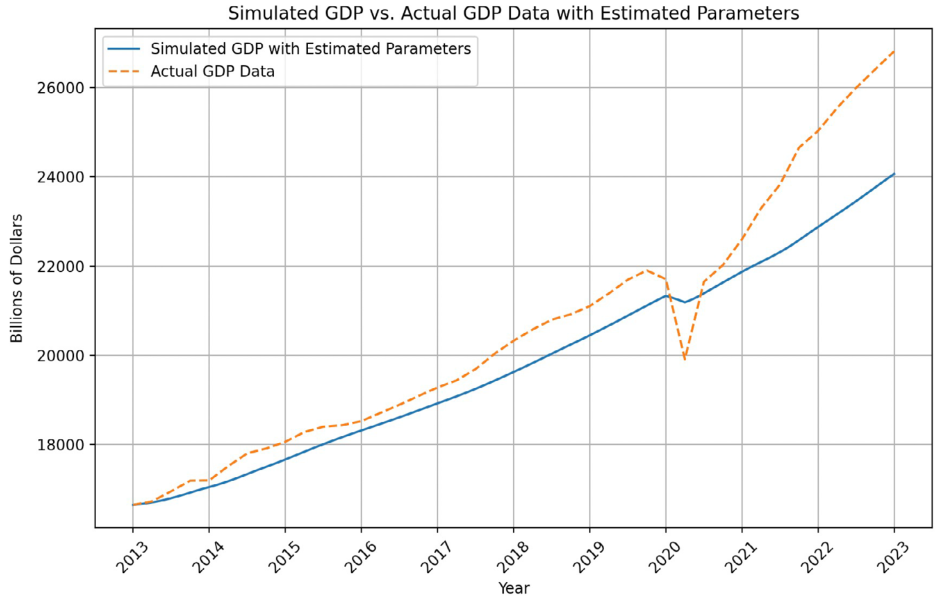
Prompt

I updated the code to include an Objective Function and goodness-of-fit metrics such as R-squared, Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Theil's bias (U). I also plotted the model predictions against actual data, which I attached to this prompt [see Figure 4]. *Here are the results*:

Estimated Parameters:
Expectation Formation Time: 1.77
Production Adjustment Time: 1.00
Alpha: 0.08

DOI: 10.1002/sdr

Fig. 4. Initial calibration results, provided in a ChatGPT prompt



Reference MPC: 0.74
Initial Expected Income: 15264.09

Statistics:
Objective Function Value: 24235675.09
R-squared: 0.85
MAPE (Mean Absolute Percentage Error): 3.39%
RMSE (Root Mean Squared Error): 1045.90
Theil's U: 0.02

Since I use the least squares method, how would I see the optimization adjustment details so that I can control the number of iterations and find the convergence criteria? I want to change them to see if I can improve the calibration results.

Response summary

ChatGPT provided guidance on adjusting the optimization settings in the *least_squares* function in our Python code. The steps include viewing optimization details by setting the verbose option to 1 or 2, which prints information about the optimization's progress to the console. To modify the number of iterations, *max_nfev* parameter, representing the maximum number of function evaluations, can be adjusted. Additionally, ChatGPT noted that the convergence criteria are defined by three parameters: *xtol*, the tolerance for changes in the solution; *ftol*, the tolerance for changes in the cost function; and *gtol*, the tolerance for the gradient. These adjustments help fine-tune the optimization process for more precise and desired outcomes in the *least_squares* function.

DOI: 10.1002/sdr

Updated optimization results

Prompt

I updated the objective function based on your suggestions, and I report the results of optimization below [we also provided optimization iteration results; see Appendix S3 in the online supporting information]. *I included the verbose option and set the maximum number of iterations to 100,000. I also set the tolerances to 1e-4 to enhance the optimization results.*

'xtol' termination condition is satisfied
Function evaluations 34, initial cost 5.4874e+08, final cost 2.9302e+06, first-order optimality 1.91e+05
Estimated Parameters:
Expectation Formation Time: 4.39
Production Adjustment Time: 0.27
Alpha: 0.28
Reference MPC: 0.84
Initial Expected Income: 13050.87
Statistics
Objective Function Value: 2930173.23
R-squared: 0.98
MAPE (Mean Absolute Percentage Error): 1.44%
RMSE (Root Mean Squared Error): 360.90
Theil's U: 0.01

I see improvements in all statistics, like R-squared has improved a lot. The overall match of the simulation results and historical trend in the attached figure has also improved [Figure 5]. *I'm happy with these new results. What do you think?*

Response summary

ChatGPT analyzed the optimization results and provided insights into their meaning. It noted that the substantial reduction in the cost (also known as payoff or objective function value) indicates a successful optimization in minimizing the discrepancy between the model and the data. Various statistics further illuminate the quality of the fit: an R-squared value of 0.98 suggests the model accounts for most variability in the response data; a Mean Absolute Percentage Error (MAPE) of 1.44% is considered excellent, showing minimal deviation from actual values; a Root Mean Squared Error (RMSE) of 360.9, while context dependent, appears small given the scale of the data; and Theil's U statistic of 0.0087, close to zero, points to a good fit. ChatGPT concluded that these metrics collectively affirm the effectiveness of the optimization process in aligning the model closely with the observed data.

Sensitivity analysis

Prompt

Now that my model is well calibrated, I extended the code to conduct a sensitivity analysis on GDP's final value by varying $\pm$ *25% the estimated values for Expectation Formation Time* (*EFT*) *and Production Adjustment Time* (*PAT*), *as*

DOI: 10.1002/sdr

Fig. 5. Improved calibration results, provided in a ChatGPT prompt

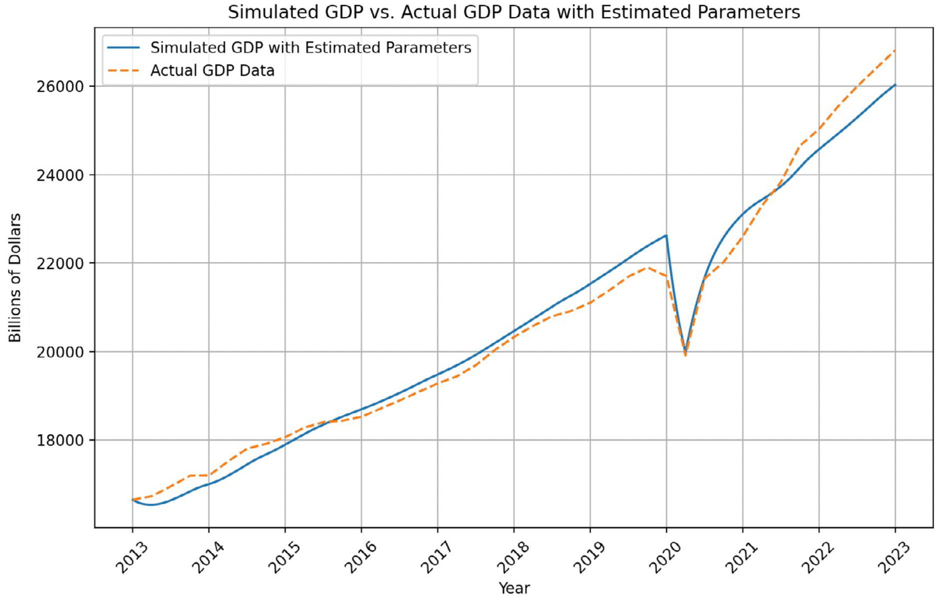
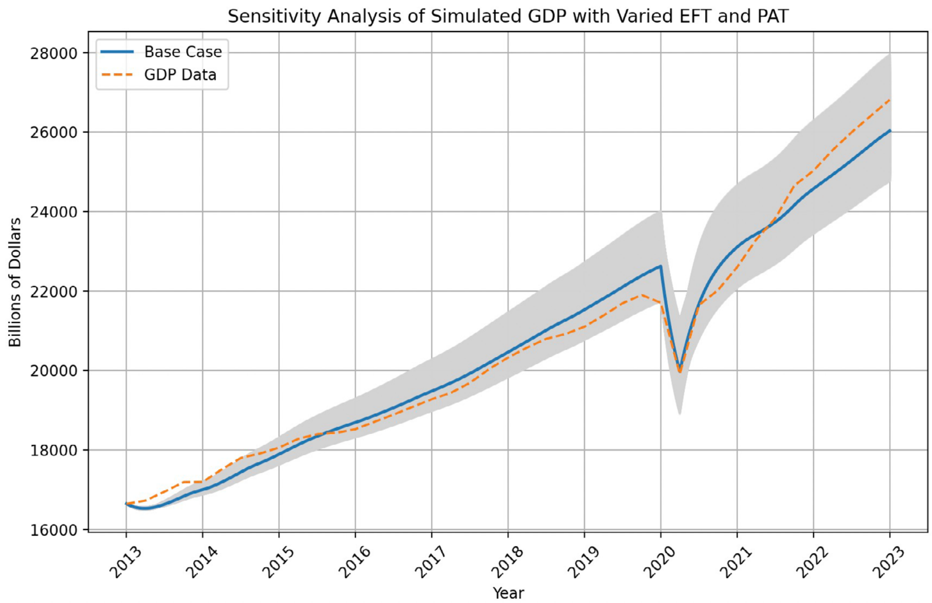


Fig. 6. Initial sensitivity analysis by varying Expectation Formation Time (EFT) and Production Adjustment Time (PAT) by ±25% of their estimated values, provided in a ChatGPT prompt



shown in the attached figure [*Figure 6*]. *My understanding is that while the results show some variations, the overall dynamic behavior of the trend doesn't change much. How does this intuition sound? Also, I want to generate a two-dimensional heatmap plot to look more closely at pairwise scenarios but I'm clueless about how to create those! Can you show me how to do so?* [We provided our code in this prompt, see Appendix S2 in the online supporting information].

DOI: 10.1002/sdr

Response summary

ChatGPT commented on our intuition about the sensitivity results. Specifically, it mentioned “When conducting a sensitivity analysis and seeing that the overall dynamic behavior of the GDP trend remains consistent despite variations in the parameters, it generally suggests that the model is robust to those parameter changes. It indicates that the economic model’s predictions are not overly sensitive to small changes in the expectation formation time (EFT) and production adjustment time (PAT), which is a good sign of model stability. This is supported by the graph you shared, where the ‘Base Case’ and ‘GDP Data’ lines follow a similar trajectory even when the shaded area, representing the uncertainty due to parameter variation, expands over time.”

In addition, ChatGPT outlined a method for producing a contour plot or heat map to analyze the impact of varying Expectation Formation Time (EFT) and Production Adjustment Time (PAT) on GDP in a given model. The suggested approach includes defining a grid of EFT and PAT values, simulating GDP for each EFT and PAT pair, storing these GDP values in a two-dimensional array aligned with the EFT and PAT grid, and then using functions like *plt.contourf* or *plt.imshow* to create a contour plot or heat map (Figure 7). This approach allows for a visual representation of the relationship between EFT, PAT, and GDP, aiding in a more intuitive understanding of the model’s dynamics [see Appendix S2 in the online supporting information for the code provided by ChatGPT].

Answering the research question

Prompt

So far, I have estimated that from 2013 to 2023, the Reference Marginal Propensity to Consume = 0.84, and the reduction due to the COVID-19 shock is 28%

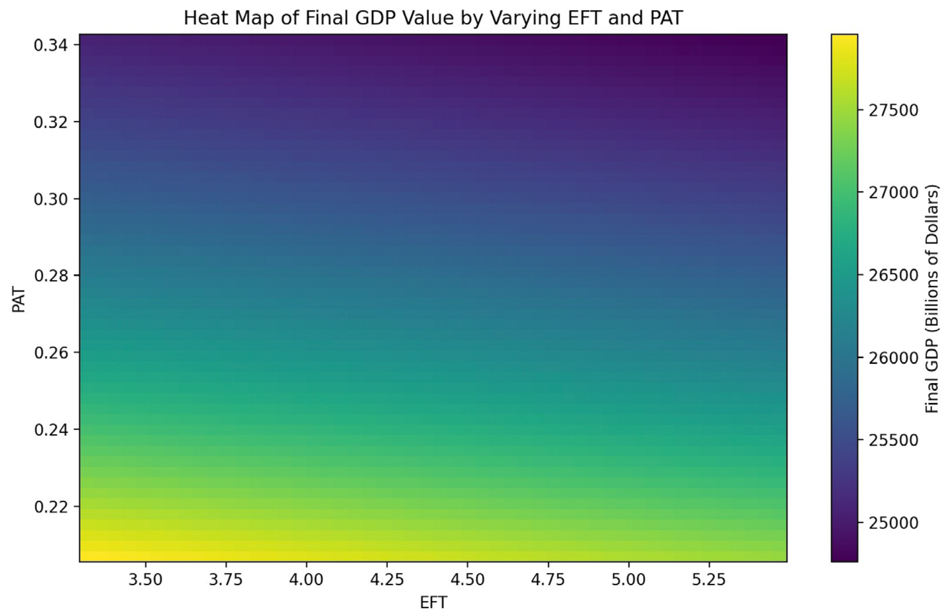


Fig. 7. Updated sensitivity graph after implementing the revised code obtained from ChatGPT

DOI: 10.1002/sdr

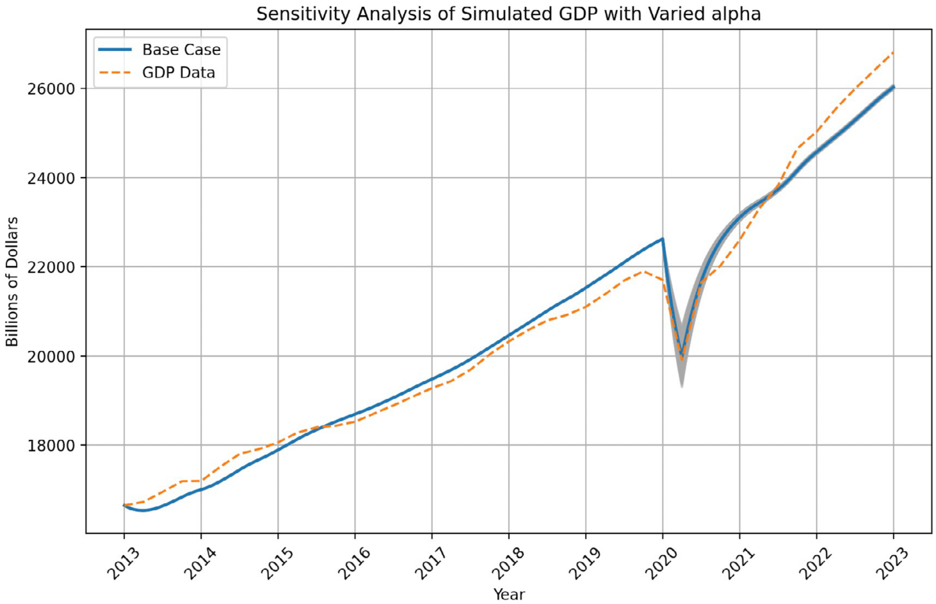


Fig. 8. Sensitivity analysis by varying percentage reduction (*alpha*) in Reference Marginal Propensity to Consume by ±25% of its estimated values, provided in a ChatGPT prompt

(*alpha = 0.28*). *I have done another analysis to observe the sensitivity of simulated GDP to ± 25% changes in the estimated value for alpha. Adjusting the estimated parameter by 25% upwards and downwards yielded consistent results, and I interpret this as a sign of the stability of my findings. Any feedback about this understanding? I have attached the figure too* [Figure 8]; *let me know what you think.*

Response summary

ChatGPT agreed with our assessment of the sensitivity analysis of simulated GDP with varying alpha values, highlighting the impact of the COVID-19 shock on the model's GDP output and that the model is not overly sensitive to alpha variations, exhibiting stability. It noted that the analysis showed that the simulated GDP is sensitive to changes in alpha, especially around the COVID-19 shock period, but converges outside of this period, indicating acute sensitivity during the shock. The actual GDP data align well within the sensitivity analysis range, suggesting that the model, despite its sensitivity to alpha, produces results consistent with observed data. Furthermore, it noted that the sensitivity analysis provides insights into potential policy implications, particularly in response to changes in alpha.

Model documentation, communication, and maintenance

Generative AI can assist in systematically documenting model development stages, ensuring every step is accurately recorded and easily understandable. Generative AI can also help communicate a simulation model and results in various forms and more accessible ways, such as developing interactive web interfaces. Finally, generative AI can aid in maintaining simulation models by

DOI: 10.1002/sdr

updating and improving the codes, for example, by translating the code into another programming language.

Model documentation

Prompt

I have my codes and the list of my model variables and parameters, along with their input, documented in a Word file [see Appendix S4 in the online supporting information; we first drafted our documentation but then used ChatGPT to revise the documentation in a separate chat]. *I think I have done a neat job in documenting my model, especially thanks to your feedback earlier that my code did not have enough explanation text. But is there anything else that I should document to facilitate the replication of my study?*

Response summary

ChatGPT reviewed our documentation and found it well-structured, encompassing key aspects crucial for understanding and replicating the simulation. However, several useful recommendations were made for further improvement: including details on model calibration to aid in accurate replication; adding a section on sensitivity analysis to highlight the robustness of the model; documenting the version control system with a link to the repository for tracking code evolution; acknowledging any known errors or limitations of the model such as potential biases or performance issues; providing case studies or application examples to illustrate practical usage; creating a user-friendly guide or tutorial for nonexperts; offering guidelines for feedback and contributions to encourage collaborative enhancements; and discussing ethical considerations and usage guidelines, especially for models with significant societal impacts.

Interactive model interface

Prompt

I'm interested in having a simple interactive interface for my model. I want to have four sliders to change the values for Reference Marginal Propensity to Consume (Reference MPC), Production Adjustment Time (PAT), Expectation Formation Time (EFT), and alpha, and the graph that be updated when I change the sliders. The graph should show Simulated GDP and Historical GDP. Of course, when I change the sliders, the Simulated GDP will be updated accordingly. Can you guide me on how to create this interface?

Response summary

ChatGPT provided guidance on creating a simple Flask application to incorporate a slider and graph feature based on an existing Python code. This process includes several steps: installing Flask using pip; writing a Python script to create a Flask application, which integrates the Ordinary Differential Equation (ODE) model and includes a function for solving the ODEs, handling HTTP requests and generating a plot; creating an HTML template with a slider and an area to display the graph; running the Flask application locally to interact with it through a web browser; and considerations for deploying the application to a web server for real-world usage, highlighting the need for handling security, managing dependencies, and potentially using a more robust web server setup. This approach

DOI: 10.1002/sdr

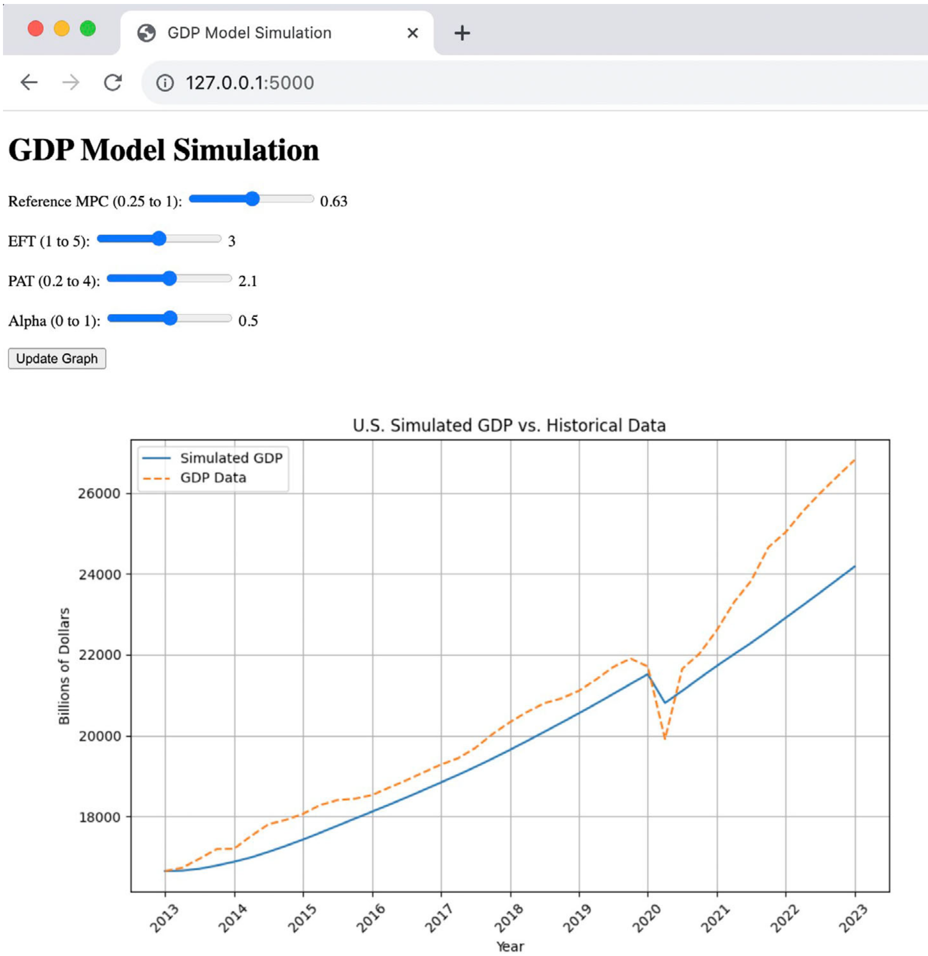


Fig. 9. Interactive dashboard created through step-by-step guideline provided by ChatGPT

provides an interactive way to visualize and manipulate the results of the ODE model through a web interface. ChatGPT's step-by-step guideline is shown in Figure S1 in the online supporting information. Following those guidelines, we were able to create the simple interface shown in Figure 9.

Translating the code from python to R

Prompt

To facilitate the replication of my model, I'd like to share the model in other languages too. My code is in Python, but can you create an R version for it?

Response summary

ChatGPT translated the Python code to R. Although it could not run R codes in the chat conversation, it provided instructions of how to run the code in R environment.

DOI: 10.1002/sdr

Considerations for using generative AI in simulation modeling

We showcased an example of using ChatGPT in a modeling exercise. However, it is essential to discuss how to use these generative AI tools effectively. Our key recommendations include using generative AI to augment critical thinking, ensuring clear communication, considering their iterative nature, recognizing their limitations, integrating interdisciplinary insights, emphasizing transparent documentation, and prioritizing ethical considerations.

First and foremost, it is crucial to understand that tools like ChatGPT are meant to augment, rather than supplant, human cognition and critical thinking. An overreliance on them can result in bypassing opportunities for thorough analysis and meaningful engagement with the modeling problems. Researchers need to be cautious about accepting AI-generated responses without scrutiny. It is imperative to cross-verify the information provided by AI to ensure its accuracy. After all, the nuances of building models are to refine our mental models, but relying uncritically on these tools undermines the spirit of modeling. The true value of generative AI is realized when it serves to challenge, broaden, and enhance the critical thinking of modelers, rather than acting as a substitute for it.

Another important consideration involves clear communication of user needs and intentions. If the responses from ChatGPT do not align with the desired direction, it is essential to provide feedback or refine the queries (Biswas, 2023). This iterative process of dialog and adjustment is fundamental, as the tool's efficacy is greatly enhanced by its learning capabilities, which are driven by user input. The more informed and specific the researcher's input and queries, the more tailored and useful the tool's output will be. A common pitfall we have observed is a quick dismissal of the tool due to dissatisfaction with initial responses without realizing the potential improvements that can be achieved through subsequent interactions and feedback.

It is also important to emphasize that engaging with tools like ChatGPT is inherently subjective and heuristic in nature. The effectiveness of refining outcomes from generative AI hinges on the researcher's depth of knowledge, expertise, and the specific nature of feedback they seek. This dynamic interaction requires an understanding of the tool's capabilities and limitations, as well as a clear vision of the desired outcome. It demands a careful orchestration of inputs where researchers must not only articulate their inquiries clearly but also interpret the tool's responses through a critical lens. This involves a discerning evaluation of the generated content for accuracy, relevance, and depth, continually refining the dialog to steer closer to meaningful insights. One must become adept at navigating the responses, identifying when to probe deeper, when to redirect the course of the conversation, and how to extract value even from seemingly tangential or unexpected AI-generated contributions.

In our case study, we adhered to a structured yet adaptive process. Initially, we formulated prompts and evaluated ChatGPT's responses. If a response did not address our objectives or lacked relevance, we incorporated more specific details. This iterative process, often involving two to three cycles per query, was crucial to ensure that each response from ChatGPT was not only relevant but also

DOI: 10.1002/sdr

informative. While our method did not follow a formal evaluation against predetermined criteria, the overall feedback-based refinement aligns with existing research for prompt design (Velásquez-Henao *et al.*, 2023; White *et al.*, 2023b). We provide an example of this iterative approach in Appendix S5 in the online supporting information.

Additionally, it is essential to recognize the evolving nature of generative AI technologies and their limitations in keeping pace with the latest research and data. These models are trained on datasets that, while extensive, may not include the most recent studies or emerging trends in simulation modeling. This lag in data can lead to gaps in the AI's knowledge base, making it crucial for researchers to supplement AI-generated insights with the latest findings and expert analyses. Furthermore, generative AI may not fully grasp the nuances of highly specialized or novel methodologies in simulation modeling, underscoring the need for researchers to critically assess and adapt AI-generated suggestions with up-to-date, domain-specific knowledge (Van Noorden and Webb, 2023).

In addition to these considerations, the integration of interdisciplinary expertise is vital in complex modeling scenarios. ChatGPT, while adept at providing information within its training, may not always capture the nuanced understanding that comes from specialized, domain-specific knowledge. Therefore, engaging with experts from relevant disciplines can greatly enhance the quality and applicability of simulation models. Collaborating with economists, data scientists, healthcare professionals, or environmental scientists, depending on the research focus, can provide critical insights beyond the scope of AI-generated content. This interdisciplinary approach enriches the modeling process and ensures a more holistic and accurate representation of real-world scenarios, thereby maximizing the effectiveness of the simulation research.

An additional critical aspect to consider is the importance of thorough documentation and transparency when using generative AI tools. Documenting how ChatGPT was used, the nature of the queries made, and how its responses were integrated into the research helps establish a clear audit trail (Graf and Bernardi, 2023). While the exact same results cannot be regenerated (i.e. replicability) due to the stochastic nature of generative AI, transparency is vital for reproducibility. Transparency in documenting the AI's role in particular is essential in maintaining accountability. It ensures that other researchers can understand the rationale behind AI use in certain instances and evaluate the influence of AI-generated content on the final conclusions. Such practices fortify the reliability of simulation research, aligning with the broader principles of scientific integrity and rigor.

Finally, ethical considerations are of paramount importance. Researchers must be acutely aware of issues such as bias and data privacy. For instance, there is a risk of inadvertently incorporating biased data into models that could skew results, potentially leading to misguided strategies (Guleria *et al.*, 2023). This underscores the need for vigilance in ensuring that the data and assumptions underpinning these simulations are as unbiased and representative as possible. Similarly, the risk of generating deceptive or harmful content with generative AI tools cannot be overstated. Moreover, simulation models often deal with sensitive data, making data privacy a critical concern. Researchers should be cautious about using generative AI in ways that might inadvertently expose confidential

DOI: 10.1002/sdr

information, especially in simulations involving personal data in sectors like health care or finance.

Overall, while generative AI tools are powerful in the realm of simulation research, their use must be tempered with critical thinking and ethical considerations. These notes are not just to prevent misuse but also to harness the full potential of AI as a complement to human expertise in the evolving landscape of simulation modeling research.

Discussion

In this article, we presented an example illustrating the use of generative AI tools in simulation modeling and then discussed how not to use them. Our report offers a balanced perspective, highlighting the importance of getting feedback and using generative AI as a research assistant rather than asking "to do the job." This report aims to enable researchers to make more informed decisions about incorporating these tools into their research methodologies responsibly.

We provided two contributions. Our first contribution is advancing the understanding of how generative AI can be integrated into simulation modeling. We demonstrated the practical application of generative AI in facilitating simulation modeling, particularly in areas where efficiency can be enhanced. By documenting and analyzing the use of generative AI, we showed their potential applications in augmenting human cognition in simulation modeling tasks. The article shows how generative AI can help streamline simulation modeling by refining research questions, enhancing model conceptualization, exploring data sources, and guiding formulation and implementation. It also shows AI's role in different aspects of modeling, such as model testing, analysis, optimization, sensitivity analysis, transparent documentation, and effective communication. In every step, these tools aid in expanding modelers' thinking processes by identifying common mistakes and suggesting alternative representations.

The second contribution of this research is the practical application of generative AI, demonstrated through a case study. This case study serves as an example of how generative AI can be utilized in simulation modeling, providing insights into handling complex, real-world problems. By considering the impact of COVID-19 as a factor in a consumption multiplier model, our research showcases the ability of generative AI to contribute to the modeling process in a practical way. This case study illustrates the potential of generative AI in enhancing the research process and provides a template for their application in other simulation modeling scenarios.

While we did not leverage ChatGPT's full range of data management capabilities in our case study, its role in this area is significant. ChatGPT can facilitate data cleaning and quality checks, assisting in tasks like identifying missing values, converting data types, and reformatting files. Beyond basic preparation, it aids in categorizing and tagging data, generating descriptive statistics, and scripting for complex manipulations across different programming languages. Additionally, ChatGPT holds promise in processing text data, such as analyzing interview transcripts to extract model mechanisms like variables and causal relationships to assist in constructing causal loop diagrams (Jalali and

DOI: 10.1002/sdr

Akhavan, 2024). Recent advancements in AI capabilities for processing extensive and varied text formats are opening up exciting new avenues for research.

As we observe the recent introduction of generative AI tools, it is important to note that the field of AI is rapidly evolving, with new developments and updates emerging regularly. As we concluded this article, Google introduced Gemini, a new tool with promising features that are yet to be evaluated. This dynamic landscape requires researchers to stay informed about the latest advancements and updates in generative AI technologies. Staying current ensures the effective utilization of these tools and helps identify new opportunities and approaches that could enhance simulation research. Researchers should be open to experimenting with new features and capabilities of generative AI, while also being adaptable in altering their methodologies in response to the evolving capabilities of these tools. This mindset of ongoing learning and adaptation is crucial for harnessing the full potential of AI in simulation modeling research.

Finally, as we discussed earlier, our observations indicate that people often briefly experiment with generative AI platforms, and if they do not immediately obtain their desired response, they quickly form a negative opinion. However, it is crucial to understand that these platforms are fundamentally about iterative learning and improvement. If the initial response is not what is needed, this should not be seen as a failure but rather as an opportunity to refine and enhance the prompt. Effectively interacting with generative AI platforms often requires practice and a certain degree of patience. It involves a process of trial and error, where each interaction serves as a teaching moment to guide the AI toward more relevant responses. By continuously tweaking and improving the prompts based on previous outputs, users can gradually steer the conversation in a more productive direction. This ongoing engagement and adaptation are essential for making the most out of what these sophisticated AI platforms have to offer.

This study is subject to several limitations. First, our findings heavily rely on the current capabilities and understanding of tools like ChatGPT, which are subject to change. Similarly, our study's focus on ChatGPT may limit the applicability of our findings to other generative AI tools with different capabilities, limitations, or operational frameworks. Second, while we aimed to keep our approach broad, it may not cover all possible variations in simulation modeling, which could limit its generalizability to specific contexts. Third, while we begin with model conceptualization, more work is needed for initial conceptualization, where various qualitative tools, such as system mapping and causal loop diagrams, are often used. However, a detailed examination of these approaches is beyond the scope of this article. Fourth, the inherent complexity and unpredictability of human-language processing in AI models suggest that certain nuances of the research topic might not have been fully captured or addressed. Fifth, the process of refining prompts for ChatGPT was subject to our interpretation and understanding, potentially not encompassing the full spectrum of possible query formulations. Sixth, while we revised our prompts to enhance their relevancy and ensure ChatGPT understood our requests, we did not analyze the sensitivity of ChatGPT's responses to variations in prompts.

Finally, it is important to note that the effectiveness of generative AI in dealing with entirely novel ideas has not been tested thoroughly, which could be subject to future research. Tools like ChatGPT are heavily trained on existing data.

DOI: 10.1002/sdr

Therefore, they might not provide accurate information, for example., a parameter value range, if they are not exposed to enough relevant data. However, they can still help review the logical aspects of the modeling. While we would not claim that generative AI is useless when working on novel ideas, we acknowledge that their ability to deal with new, unreviewed scientific theories is limited by the lack of available data and their inability to conduct independent verification or experimentation, an area that warrants exploration in future research.

Despite these limitations, our report represents a first step toward understanding the potential of generative AI for simulation modeling. We would like to emphasize that our report demonstrates how these tools can be iteratively employed at every phase of simulation modeling, not only to facilitate but also to augment critical thinking and overall research methodology. Our aim is to empower researchers with critical thinking to utilize the best available tools for idea development and content review in their simulation studies. By highlighting both the possibilities and the challenges of employing generative AI in simulation modeling, we hope to encourage other researchers in further exploration and innovation in this evolving field.

Acknowledgement

The authors thank Satoshi Koiso, Hannah Lee, Hesam Mahmoudi, Elizabeth Mason, Amir Salarkia, and Melike Yildirim, who provided constructive feedback on earlier versions of this article.

Funding information

MSJ was supported in part by grant R01CE003358 from the U.S. Centers for Disease Control and Prevention.

References

Ariyaratne S, Iyengar KP, Botchu R. 2023. Will collaborative publishing with ChatGPT drive academic writing in the future? *British Journal of Surgery* **110**(9): 1213–1214. https://doi.org/10.1093/bjs/znad198.

Banks J. 2005. *Discrete Event System Simulation*. Upper Saddle River, NJ: Pearson Prentice Hall.

Biswas SS. 2023. ChatGPT for research and publication: a step-by-step guide. *The Journal of Pediatric Pharmacology and Therapeutics* **28**(6): 576–584. https://doi.org/10.5863/1551-6776-28.6.576.

Castellanos-Gomez A. 2023. Good practices for scientific article writing with ChatGPT and other artificial intelligence language models. *Nano* **3**(2): 135–138. https://doi.org/10.3390/nanomanufacturing3020009.

Davis JP, Eisenhardt KM, Bingham CB. 2007. Developing theory through simulation methods. *Academy of Management Review* **32**(2): 480–499.

Dowling M, Lucey B. 2023. ChatGPT for (finance) research: the Bananarama conjecture. *Finance Research Letters* **53**: 103662. https://doi.org/10.1016/j.frl.2023.103662.

DOI: 10.1002/sdr

Flanagin A, Bibbins-Domingo K, Berkwits M, Christiansen SL. 2023. Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. *JAMA* **329**(8): 637–639. https://doi.org/10.1001/jama.2023.1344.

Ghaffarzadegan N, Majumdar A, Williams R, Hosseinichimeh N. 2024. Generative agent-based modeling: an introduction and tutorial. *System Dynamics Review* **40**(1): e1761. https://doi.org/10.1002/sdr.1761.

Graf A, Bernardi RE. 2023. ChatGPT in research: balancing ethics, transparency and advancement. *Neuroscience* **515**: 71–73. https://doi.org/10.1016/j.neuroscience.2023.02.008.

Guleria A, Krishan K, Sharma V, Kanchan T. 2023. ChatGPT: Ethical concerns and challenges in academics and research. *Journal of Infection in Developing Countries* **17**(9): 1292–1299. https://doi.org/10.3855/jidc.18738.

Harrison JR, Lin Z, Carroll GR, Carley KM. 2007. Simulation modeling in organizational and management research. *Academy of Management Review* **32**(4): 1229–1245. https://doi.org/10.5465/amr.2007.26586485.

Homer JB. 1996. Why we iterate: scientific modeling in theory and practice. *System Dynamics Review* **12**(1): 1–19. https://doi.org/10.1002/(SICI)1099-1727(199621)12:1<1::AID-SDR93>3.0.CO;2-P.

Jalali MS, Akhavan A. 2024. *Integrating AI language models in qualitative research: Replicating interview data analysis with ChatGPT*. *System Dynamics Review* **40**(3): e1772. https://doi.org/10.1002/sdr.1772

Jalali MS, Beaulieu E. 2023. Strengthening a weak link: Transparency of causal loop diagrams—Current state and recommendations. *System Dynamics Review*. https://doi.org/10.1002/sdr.1753.

Jalali MS, DiGennaro C, Guitar A, Lew K, Rahmandad H. 2021. Evolution and reproducibility of simulation modeling in epidemiology and health policy over half a century. *Epidemiologic Reviews* **43**(1): 166–175. https://doi.org/10.1093/epirev/mxab006.

Law AM, Kelton WD. 2007. *Simulation Modeling and Analysis* (4th edn). New York: Mcgraw-Hill.

Low G. 1980. The multiplier-accelerator model of business cycles interpreted from a system dynamics perspective. In Randers J (ed.), *Elements of the system Dynamics Method*. MIT Press: Cambridge; 76–94.

Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng P-C, Bright TJ, Tatonetti N, Won KJ, Gonzalez-Hernandez G *et al*. 2023. ChatGPT and large language models in academia: opportunities and challenges. *Biodata Mining* **16**(1): 20. https://doi.org/10.1186/s13040-023-00339-9.

Monks T, Currie CSM, Onggo BS, Robinson S, Kunc M, Taylor SJE. 2019. Strengthening the reporting of empirical simulation studies: introducing the STRESS guidelines. *Journal of Simulation* **13**(1): 55–67. https://doi.org/10.1080/17477778.2018.1442155.

Nature Editorial. 2023. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature* **613**(7945): 612. https://doi.org/10.1038/d41586-023-00191-1.

Rahmandad H, Sterman JD. 2012. Reporting guidelines for simulation-based research in social sciences. *System Dynamics Review* **28**(4): 396–411. https://doi.org/10.1002/sdr.1481.

Samuelson PA. 1939. Interactions between the multiplier analysis and the principle of acceleration. *The Review of Economics and Statistics* **21**(2): 75–78. https://doi.org/10.2307/1927758.

Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, Moy L. 2023. ChatGPT and other large language models are double-edged swords. *Radiology* **307**(2): e230163. https://doi.org/10.1148/radiol.230163.

DOI: 10.1002/sdr

Sterman JD. 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Irwin/McGraw-Hill: Boston.

Stokel-Walker C, Van Noorden R. 2023. What ChatGPT and generative AI mean for science. *Nature* **614**(7947): 214–216. https://doi.org/10.1038/d41586-023-00340-6.

Thapa S, Adhikari S. 2023. ChatGPT, bard, and large language models for biomedical research: Opportunities and pitfalls. *Annals of Biomedical Engineering* **51**(12):2647–2651. https://doi.org/10.1007/s10439-023-03284-0.

Thorp HH. 2023. ChatGPT is fun, but not an author. *Science* **379**(6630): 313. https://doi.org/10.1126/science.adg7879.

Van Noorden R, Webb R. 2023. ChatGPT and science: the AI system was a force in 2023—For good and bad. *Nature* **624**:509. https://doi.org/10.1038/d41586-023-03930-6.

Velásquez-Henao JD, Franco-Cardona CJ, Cadavid-Higuita L. 2023. Prompt engineering: a methodology for optimizing interactions with AI-language models in the field of engineering. *DYNA* **90**(230): 9–17. https://doi.org/10.15446/dyna.v90n230.111700.

Wang X, Sanders HM, Liu Y, Seang K, Tran BX, Atanasov AG, Qiu Y, Tang S, Car J, Wang YX *et al*. 2023. ChatGPT: promise and challenges for deployment in low- and middle-income countries. *The Lancet Regional Health Western Pacific* **41**. https://doi.org/10.1016/j.lanwpc.2023.100905.

White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, Elnashar A, Spencer-Smith J, Schmidt DC 2023a. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint*. https://doi.org/10.48550/arXiv.2302.11382.

White J, Hays S, Fu Q, Spencer-Smith J, Schmidt DC 2023b. ChatGPT prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. *arXiv preprint*. https://doi.org/10.48550/arXiv.2303.07839.

Xiang L, Tang M, Yin Z, Zheng M, Lu S. 2021. The COVID-19 pandemic and economic growth: theory and simulation. *Frontiers in Public Health* **9**. https://doi.org/10.3389/fpubh.2021.741525.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's website.

Appendix S1. Generative AI and Simulation Modeling: How Should You (Not) Use Large Language Models Like ChatGPT.

DOI: 10.1002/sdr