Performance of Large Language Models in Graduate-Level System Dynamics Exams

Burak Çetiner[1,2], Gönenç Yücel[2] and Yaman Barlas[2,3]

[1]Marmara University, [2]Boğaziçi University, [3]Acıbadem University

Abstract:

Large language models (LLMs) are attracting growing attention in System Dynamics (SD) research and practice, particularly for tasks such as conceptual modeling, parameter calibration, and interpretation of simulation outputs. On the other hand, there is very limited evidence on the performance of LLMS in core SD problem-solving. This study evaluates the current capabilities of three widely used LLMs through an exam-based benchmarking in graduate-level SD education. Specifically, the models were asked to answer real midterm and final examination questions under a standardized, minimal prompt setting intended to provide an estimate of performance. Their responses were graded by the course instructor and compared directly with the performance of the student cohort. The findings indicate that ChatGPT 5.2 and Gemini 3 performed above the class average on both examinations, while DeepSeek performed closer to the class average overall. Across question types, the models were relatively strong in short answer questions, stock-flow identification, equation formulation, and qualitative interpretation of dynamic behavior, but weaker in tasks requiring explicit graphical representation of dynamic trajectories. By comparing LLMs with human graduate students in an authentic course setting, this study hopes to contribute to benchmarks for assessing LLM competence in SD. The results suggest that contemporary LLMs can already demonstrate substantial proficiency in several foundational SD skills, while also revealing persistent limitations in formal and visual dynamic representation. These findings can contribute to the formation of an empirical baseline for future research on prompting strategies and the effective integration of LLMs into SD education and modeling practice.

1. Introduction

In recent years, data science has been developing very quickly and achieving great accomplishments. Starting from machine learning algorithms, the field has exhibited important progress and practical applications for the various fields of technology. The rise of Large Language Models (LLM) made a huge leap in order to improve artificial intelligence (AI) studies. Vaswani et al. (2017) introduce a working mechanism called “Attention” and this mechanism enables language models to process very large amounts of data in a computation extensive process. It was such a revolutionary idea and changed artificial intelligence studies forever.

On the other hand, the field of System Dynamics (SD) is improving in such a direction that there is an increase in data-driven methodologies. The related literature will be briefly reviewed below in the literature review chapter.

This study examines the performance of contemporary LLMs in solving System Dynamics problems. To this end, we develop an experimental design in which three widely used LLMs (ChatGPT, Gemini, and DeepSeek) are tasked with answering authentic midterm and final examination questions from the course IE 550: *Dynamics of Socio-Economic Systems*, offered in the Industrial Engineering Department at Boğaziçi University. The LLM's responses are then analyzed to assess their problem-solving capabilities within the context of System Dynamics competencies. The LLM generated responses are compared both against students' answers and across the LLMs themselves. In addition, exam scores are statistically analyzed to evaluate whether each model's performance exceeds that of an average graduate student enrolled in the course.

2. Aim of the Study

In this study, we aim to systematically assess the capabilities and limitations of widely used LLMs in SD problem solving. Understanding LLM performance in the domain of System Dynamics can inform their efficient and effective use in research and practice. On the other hand, identifying where these models fall short can help clarify the underlying sources of error and guide targeted mitigation strategies. As a brief result, these analyses can contribute to methodological advances and improved integration at the intersection of AI, machine learning and System Dynamics. There is a quite recent literature on the use of LLMs in System Dynamics problems and applications. The researchers use LLMs for conceptual modelling phase, parameter calibration, and output analysis. In this research we follow different approach in order to asses LLMs performance in System Dynamics problem solving in an education setting. To this end, we treat each LLM like a graduate student who takes the same midterm and final exams that regular students take.

3. Literature Review

The integration of Artificial Intelligence (AI) and Machine Learning (ML) into System Dynamics has evolved significantly, transitioning from quantitative calibration to qualitative conceptualization. Historically, AI and ML algorithms were primarily utilized for parameter estimation, model calibration, output analysis, and sensitivity analysis. The pattern recognition capabilities of machine learning algorithms offered substantial accuracy improvements (Gadewadikar and Marshall, 2024). This alignment with data-driven approaches was anticipated by foundational figures in the field; Forrester (1980) and Sterman (2018) emphasized the critical role of data, new methods and future technologies in enhancing SD. The emergence of Large Language Models has

unlocked the potential for automating the conceptual modeling phase, namely, the construction of Causal Loop Diagrams and Stock and Flow Diagrams.

Current research demonstrates that LLMs can now facilitate the entire model building process through methods such as "text-to-model" generation, fundamentally altering the landscape of SD (Liu & Keith, 2024; Hosseinichimeh et al., 2024). A recent systematic literature review by Muthiah et al. (2025) highlights that while this integration is currently growing, it has already yielded promising results, particularly in system conceptualization. Early studies report that LLM-powered tools, such as the "System Dynamics Bot," can identify approximately 60% of causal links and feedback loops from textual data (Hosseinichimeh et al., 2024). Despite these successes, significant gaps remain. The literature indicates a negative correlation between model complexity and LLM accuracy; as models become more intricate, the reliability of AI-generated structures diminishes (Schoenberg et al., 2025). Furthermore, existing studies often lack fully automated end-to-end pipelines and robust verification mechanisms to prevent "hallucinations" or logical inconsistencies in the generated models (Muthiah et al., 2025; Wang et al., 2025). LLMs are moving from passive research assistants to active participants in the modeling process, capable of simulating complex human behaviors and identifying causal structures that are often ambiguous to the human mind. Yet, the persistent need for human oversight and the challenges of "black box" interpretability remain central themes in the ongoing discourse. (Akhavan and Jalali, 2024)

4. Design of Experiment

In this experiment, we want to investigate the current performance levels of Large Language Models (LLMs) in System Dynamics field. To this end, we evaluate three models: ChatGPT 5.2, Gemini 3, and DeepSeek. We use the professional versions of ChatGPT 5.2 and Gemini 3, and we run all models in their “deep thinking” mode to approximate their highest available reasoning capacity.

The experimental materials are the exam questions of the course IE 550 *Dynamics of Socio-Economic Systems*, offered by the Industrial Engineering Department at Boğaziçi University. The LLMs respond to authentic midterm and final examination questions provided by the course instructor (Gönenç Yücel) from the Fall 2024 semester. The cohort size in that offering was 28 students, which allows for statistically meaningful comparisons between the models’ performance and students’ exam responses.

To control for the effects of prompting strategies, we provide each model only with the PDF file containing the exam questions and no additional written guidance, except for the single instruction, “give the answers of the exam questions,” which is included solely to ensure that the models respond to the questions in English. This design is intended to estimate a lower-

bound level of LLM performance under standardized conditions. We note that more elaborate prompting strategies could plausibly yield higher performance for each model.

The answers given by the LLMs are graded by the instructor of the course, Assoc. Prof. Gönenç Yücel. Each LLM is treated like a student in the class and the evaluation is done at the level of individual questions. The answers and grades of LLMs are both compared to the students' answers and grades as well as each other's.

The questions from both the midterm and final examinations are provided in the Appendix. In the midterm examination, the first question consists of short-answer questions requiring verbal reasoning about the key concepts of System Dynamics. The second question focuses on graphical integration, while the third is a modelling question. The fourth question asks students to identify the possible dynamic behavior of a system, and the fifth requires the formulation of a stock-and-flow diagram, including time delays. Similarly, in the final examination, the first question consists of short-answer items involving verbal reasoning. The second question addresses time delays and the expected dynamic behavior of the system, while the third requires critique of equation formulation. The fourth and fifth questions are modelling questions, both involving time delays.

5. Experimental Results

The exam results are summarized in the Table 1 below. Considering the midterm exam, ChatGPT 5.2 and Gemini 3 achieved scores of 78 and 76, respectively, both of which are notably above the class average (67.13). In contrast, DeepSeek scored 60, which is below the class average. But it is also important to note that the student standard deviation is quite high (19.31), so it is very difficult to reach a 'statistical significance' conclusion with this limited data. Ranking all examinees including the LLMs by midterm performance in descending order, ChatGPT 5.2 placed 11th, Gemini 3 placed 13th, and DeepSeek placed 24th in a combined cohort of 31.

For the final exam, the LLMs exhibited stronger overall performance. ChatGPT 5.2 and Gemini 3 obtained comparable scores 72 and 76, respectively, while DeepSeek scored 67. All three scores exceeded the class average for the final which is 58.60 with a standard deviation of 21.12. These results indicate that the models performed well relative to the distribution of student outcomes, but it is impossible to do carry out a statistical significance analysis without more exams to estimate the standard deviation of LLM performances. Ranking all examinees including the LLMs by final exam performance in descending order, ChatGPT 5.2 placed 9th, Gemini 3 placed 8th, and DeepSeek placed 12th in a cohort of 31.

Table 1. Exam Grades

Exam Grades Table	Chatgpt 5.2	Gemini 3	Deepseek	Class Average (without LLMs)	Class Standard Deviation
Midterm Exam	78	76	60	67.13	19.31
Final Exam	72	76	67	58.60	21.12

These results suggest that ChatGPT 5.2 and Gemini 3 perform better in a graduate-level System Dynamics course than the average student. Although DeepSeek does not perform as strongly as the other two models, its overall performance can still be considered close to the class average when both the midterm and final examinations are taken into account.

Table 2. Question-Level Exam Grades

	Midterm Exam					Final Exam				
	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**
Chatgpt 5.2	100	80	60	75	76	90	70	86.67	55	64
Gemini 3	100	60	35	100	76	70	80	93.33	60	80
Deepseek	90	13.33	45	45	64	95	65	86.67	45	52
Class Average	**74.11**	**90.95**	**51.61**	**67.68**	**57.57**	**59.26**	**79.95**	**34.07**	**52.78**	**63.85**

Table 3. Overall Grades of LLMs

	Midterm Exam		Final Exam		Semester Grade
	Exam Grade	**Standard Deviation of Question Grades**	**Exam Grade**	**Standard Deviation of Questions Grades**	**%60 of Midterm Exam + %40 of Final Exam**
Chatgpt 5.2	78	14.35	72	14.91	75.6
Gemini 3	76	27.72	76	12.47	76
Deepseek	60	18.94	67	21.62	62.8
Class Average	67.13	21.34	58.60	18.59	63.72

The performance of the LLMs was also evaluated at the level of individual questions at Table 2. According to Table 2, all 3 LLMs perform worse in the second question of the both exams (except the Gemini's grade in the final exam) in which it is asked to draw the stock variable's behavior given the inflow and outflows. The scores lower than the class averages are marked by red in the tables. Deepseek has some low grades compared to class average in question 3, 4, and 5. However, its performance on verbal questions provides a total grade which is close to class average (60 and 67 respectively).

Overall Grades of LLMs are given in Table 3. The standard deviations of the LLMs' question level grades are mostly lower than the students' results except the midterm exam of Gemini and final exam of Deepseek. We can comment that the LLMs' performance may be more consistent, having lower standard deviations within questions compared to other students.

This analysis indicates that the models are particularly successful in short answer questions. They are generally able to identify stocks and flows correctly and to formulate stock equations with minor errors. In addition, they can often infer the qualitative behavior of a system from the coefficients or parameters governing the inflows and outflows, although DeepSeek shows greater difficulty in capturing these dynamics accurately.

However, all three LLMs lost significant points on questions requiring students to draw the dynamic trajectory of a stock variable based on given flow graphs. Rather than explicitly producing the requested graph, the models tended to provide verbal descriptions of its likely shape, which constituted inadequate answers. Nevertheless, these descriptions were, in most cases, conceptually reasonable. This suggests that, with more effective prompting strategies, LLMs may be able to generate the required graphical representations more successfully. At this point, we have to remind that this experimental design is intended to estimate a lower-bound level of LLM performance under standardized conditions. We should note that more elaborate prompting strategies could plausibly yield higher performance for each model.

6. Contribution and Conclusions

This paper aims to contribute to the literature by offering a novel way to evaluate the capabilities of Large Language Models in the field of System Dynamics. While existing studies mainly examine the use of LLMs in specific SD tasks such as conceptual modeling, causal loop diagram generation, parameter calibration, or output analysis, this study evaluates LLM's SD capabilities more generally, by treating them as if they were graduate students taking authentic course examinations. Using real midterm and final exam questions from a graduate-level System Dynamics course should provide a more realistic and academically meaningful benchmark than task-specific or artificial test settings.

A second contribution of the paper is that it compares LLM performance directly with human student performance rather than assessing the AI tools in isolation. The study positions their scores relative to class averages, score distributions, question level assessment, and ranking within a real cohort. This makes the findings easier to interpret in practical terms and shows how close current LLMs can be to graduate level student competence in System Dynamics. In this sense, the paper provides empirical evidence that some contemporary LLMs can perform at or above the average student level in SD related tests and exams.

This study examined the current performance of contemporary Large Language Models in the field of System Dynamics by evaluating ChatGPT 5.2, Gemini 3, and DeepSeek on authentic graduate level midterm and final examination questions. The results show that LLMs may have already reached a meaningful level of competence in several core System Dynamics skills. In particular, ChatGPT 5.2 and Gemini 3 performed above the class average in both exams, while DeepSeek produced results that were closer to the class average overall, especially when both exams are considered together. These findings suggest that current LLMs are capable of answering many System Dynamics questions at a level comparable to graduate students, especially in short answer tasks, stock-flow identification, equation formulation, and qualitative interpretation of system behavior.

At the same time, the findings reveal important limitations that prevent these models from being considered fully reliable substitutes for human reasoning in System Dynamics. All three LLMs showed weaknesses in questions requiring graphical representation of dynamic behavior, often describing the expected pattern verbally rather than producing the requested diagram. This indicates that although LLMs can demonstrate substantial conceptual understanding, they still face difficulties in translating that understanding into precise visual or formal outputs. Since the models were tested under a standardized and minimally prompted setting, the study provides a lower-bound estimate of their current performance.

7. Ongoing and Future Work

In this study, we only used one midterm and one final exam and asses the performance of the LLMs both in question level and cumulative level. However, to conduct a reasonable statistical analysis, the sample size should be increased in terms of examination. We are currently collecting data by increasing the examination number and recording the answers of LLMs. The next step of this study will be examining at least three midterm and three final exams and conducting statistical analysis by using sample means and standard deviation. We are also planning to analyze these results in question level to make inferences about LLMs' capabilities and limitations.

Future research can extend this study in several important directions. First, the same evaluation framework can be applied to a broader range of System Dynamics courses, institutions, and

question types in order to assess the generalizability of the findings beyond a single graduate course. Second, future studies should examine how different prompting strategies, multimodal inputs, and tool-supported environments affect LLM performance, especially on tasks involving graph drawing, causal reasoning, and the construction of stock and flow diagrams. It would also be valuable to evaluate newer model versions and compare their progress over time. This experiment design can be extended to other commonly used LLMs such as Claude and Llama. Finally, future work should move beyond exam performance alone and investigate how LLMs can be integrated into actual System Dynamics practice, such as conceptual modeling, policy design, model validation, and simulation-based analysis, while also addressing issues of reliability, interpretability, and hallucination.

Appendix A

Midterm Exam

1. (20 pts) Give short answers to the following questions:

a. Can a single-stock model endogenously exhibit overshoot-then-collapse behavior? Explain briefly.

b. What does a 'proper dt value' mean and why is it important in simulation?

c. Consider the boar hunting example that is modeled in an earlier lab session. So far we discussed a set of potential improvements to that model. List two of such potential improvements in the model structure.

d. We discussed two sorts of delays with very different natures in the context of the corporate growth model. Discuss their main conceptual differences, and point out potential locations of such delays in the discussed model structure.

e. Consider the following pairs of variables. Determine which pairs represent a direct causal-effect relationship and which are purely statistical correlations (Explain briefly)

i. Defective products shipped - Level of finished goods inspection

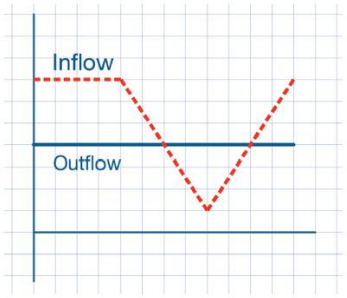
ii. Bribery incidences - Level of inflation rate

iii. Fatigue - Overwork

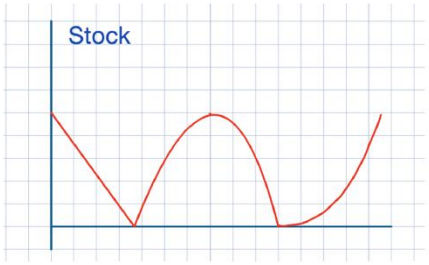
iv. Room temperature - Radiator output

2. (15 pts) Answer the following questions via plotting the asked behavior of the specified variable on the right-hand side

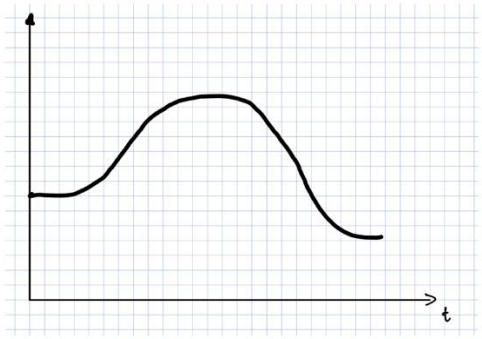
a. Given the following outflow and inflow dynamics associated with a stock variable, draw roughly the behavior patterns exhibited by the stock variable



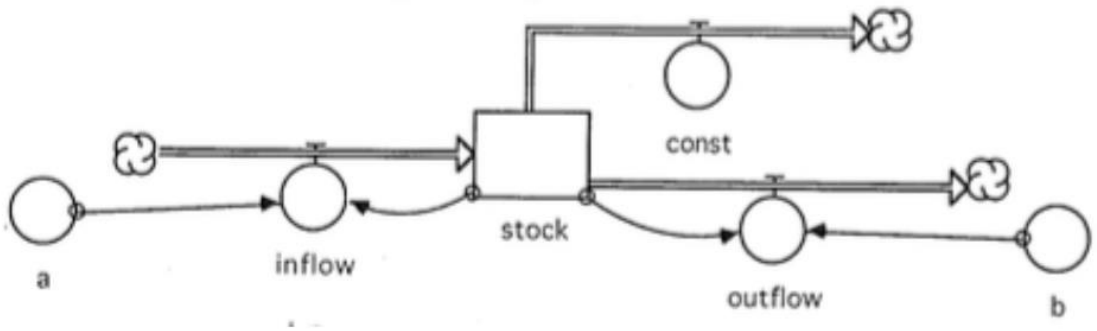
b. Given the following dynamic pattern exhibited by a stock variable, plot the corresponding "net rate of change" graph.



3. (20 pts) Consider a simple workforce hiring/firing structure in which the purpose is to bring the workforce stock to a "desired" level (which is constant in this simple goal-seeking loop).
a. Provide the stock-flow structure and equations for the variables in the diagram.
b. Plot what dynamic behavior you expect from this model.
c. Assume that the desired workforce level is also a dynamic variable and dependent on other system variables. Roughly plot the expected hiring/firing rates as well as the workforce level, if the desired workforce changes as follows. You can assume that workforce is at its desired level at t=0



4. (20 pts) In the following model, "inflow and outflow" are simply proportional to the stock and **const** is another constant outflow. Assuming that a is greater than or equal to b and both positive constants, what are the possible dynamic behaviors?



5. (25 pts) Consider a very simple model of forest regeneration, growth and death. The trees in the forest are categorized as young and mature trees. What distinguished these two is their ability to produce seed. While the mature trees can generate seeds, young ones cannot. Every year 75% of the mature trees produce seeds and spread them around. On average, a mature tree produces 800 seeds. However, not all those seeds grow, and turn into trees. The chances of a seed to grow is known to be 1/10000. Once a seed grows, it is considered as a new young tree in the forest. ON average, it takes 10 years for a tree before it can start producing seeds (hence, considered as a mature tree). Once they become mature, trees have an average life expectancy of 20 years.

a. Draw the causal-loop diagram of the described system
b. Draw the stock-diagram of the system
c. Can you identify a long-run equilibrium state for this forest?
d. Assuming that the initial stock values are YoungTrees(0)=1000 and MatureTrees(0)=3000 and a continuous model, plot roughly what behavior you expect in the long-term.

Final Exam

1. (20 pts) Answer the following questions briefly

a. In using graphical functional relations, why is "normalizing" the input variable is important? Why is proper selection of "input range" important? You can explain on a small example

b. State in what situations you would prefer a multiplicative effect formulation and in what situation you would prefer an additive effect formulation.

c. Discuss the relative importance of structural validation in a correlational forecasting model versus a causal-descriptive system dynamics model

d. Give an example for a decision formulation that would violate the so-called Baker criterion

2. (20 pts) Answer the following questions in the given sequence.

a. Consider a general 2nd order material delay (with input IN, the delayed output OUT and delay constant DEL). Give the stock-flow diagram and equations.

b. Now assume that the output OUT of this material delay structure creates an input (INFO_IN) to a 1st order information delay structure (again with a delay time of DEL). Build a stock-flow diagram of this 'hybrid' delay structure and complete the equations.

c. Compute the equilibrium levels of the variables, assuming the input IN is constant and equal to P.

d. Now assume that the system was initially at equilibrium for some duration, and then the input IN of the system is suddenly increased to a new (disequilibrium) value. Plot the expected dynamics of the final output of the entire structure.

3. (15 pts) Below you are given a set of formulations from different models. Critique these formulations, stating if they have any specific violations of the 'principles of good formulation'. Next suggest your own improved version. If needed, define your parameters and variables as necessary; but your answer must be focused and specific, do NOT extend the formulations to include factors that are not specified/mentioned.

a. A company's advertising spendings are mainly driven by the advertising activity of its competitor. Desired level of the weekly advertising spending (TL/wk) is formulated as

DesAdv = MIN(AdvertisingBudget, k*(ComptetitorAdvertising))

b. Total grazing rate of a sheep herd on open grass depends on grass availability and the sheep population; and the following formulation is proposed;
$GrassConsumption = k_1*Sheep*Grass + k_2*Grass^{0.5}$

c. In an urban population model, the in-migration and out-migration rates were aggregated together into a net migration rate NMIG. NMIG can be ≥ 0 or ≤ 0 depending on whether more people enter than leave the city.
NMIG = NMF*Pop*EJ*EH
where:
NMIG = net migration (people/year)
EJ = effect of job market on net migration (dimensionless)
EH = effect of housing market on net migration (dimensionless)
NMF = normal migration fraction (fraction/year)
Assume that EJ and EH are well-formulated graphical functions.

4. (20 pts) Weekly demand for a product is believed to be mainly influenced by two factors: Quality and Price of the product. The effect of Quality is significantly delayed (after an average of 4 weeks). We also know that a 2nd order continuous delay is suitable to model the situation. Effect of price is simpler, we assume that there is no significant time delay before the price influences the demand.

Give a S-F model sub-structure, and the complete set of equations and functions to represent the above formulation. Define your own parameters as necessary.

5. (25 pts) A bear population is known to live in the mountains of a certain region. Naturally, the bear population has a natural limit imposed by food and shelter availability (Capacity). Up to this limit, the bear population demonstrates a positive rate of change, and beyond this limit net rate of change becomes negative. The natural park authority, which monitors the region as well as the bear population issues bear hunting licenses as a means to control the bear population. The authority has a target bear population (Target), and when the perceived population goes above this target, issues new permits. The number of permits to be issued is linearly dependent on the difference between the target population and the perceived population. The authority perceives the changes in the bear population only after a delay (MonitorDelay), which can be captured as a simple perception delay. The permits have a validity period, and they expire after a delay of two years.

The number of bears hunted depends on the number of licenses. The main factor that influences the number of bears hunted per license (HuntPerLicense) is the bear density in the region (Population/Capacity). The bear density has a non-linear effect on the hunted animals per license.

Formulate the above described model. Give the stock-flow diagram corresponding to the parts above. Give all related equations and necessary functions and define your own parameters and variables when necessary.

References

- Akhavan, A., and M. S. Jalali. 2023. “Generative AI and SimulationModeling: How Should You (Not) use Large Language Models LikeChatGPT.” *System Dynamics Review* 40, no. 3: e1773.
- Barlas, Y. 2007. System dynamics: systemic feedback modeling for policy analysis. *System*, *1*(59), 1-68.
- Forrester, JW. 1970. Urban dynamics*.; Industrial Management Review 11*(3): 67.
- Forrester, JW. 1980. Information sources for modeling the national economy. *Journal of the American Statistical Association* 75(371): 555-574.
- Gadewadikar J, Marshall J. 2024. A methodology for parameter estimation in system dynamics models using artificial intelligence. *Systems Engineering* 27(2): 253–266.
- Hosseinichimeh N, Majumdar A, Williams R, Ghaffarzadegan N. 2024. From text to map: a system dynamics bot for constructing causal loop diagrams. *System Dynamics Review* 40(3): e1782.
- Jalali, M. S., and A. Akhavan. 2024. “Integrating AI Language Modelsin Qualitative Research: Replicating Interview Data Analysis WithChatGPT.” *System Dynamics Review* 40: e1772
- Liu NYG, Keith DR. 2025. Leveraging large language models for automated causal loop diagram generation: enhancing system dynamics modeling through curated prompting techniques. arXiv 2503.21798
- Muthiah AD, Turan H, El Sawah S. 2025. AI for system dynamics: mapping progress across six modelling stages. *Proceedings of the System Dynamics Society*.
- Schoenberg W, Girard D, Chung S, O'Neill E, Velasquez J, Metcalf S. 2025. How well can AI build SD models? *arXiv* 2503.15580.
- Sterman JD. 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World.* Irwin/McGraw-Hill, Boston.
- Sterman JD. 2018. System dynamics at sixty: the path forward. *System Dynamics Review* 34(1-2): 5–47.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, AN., Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30: 5998-6008.
- Veldhuis, G. A., D. Blok, M. H. T. de Boer, G. J. Kalkman, R.M. Bakker, and R. P. M. van Waas. 2024. “From Text to Model: Leveraging Natural Language Processing for System Dynamics Model Development.” System Dynamics Review 40: e178
- Wang Q, Wu J, Tang Z, Luo B, Chen N, Chen W, He B. 2025. What limits LLM-based human simulation: LLMs or our design? *arXiv* 2501.08579.