Exploring the use of Large Language Models with Chain-of-thought Prompting as an Aid in Causal Loop Diagram Development

Ashish Kumar[1,2*], Adam Quek[1,2], Michael Dorosan[1,2], Sean Shao Wei Lam[1,2,3**]

[1]SingHealth Health Services Research Centre, Singapore Health Services, SINGAPORE

[2]Health Services & Systems Research, Duke-NUS Medical School, SINGAPORE

[3]School of Computing and Information Science, Singapore Management University, SINGAPORE

*Joint first authors

** Corresponding author. Email gmslasws@nus.edu.sg

Extended Abstract

Background

The rapid development of Generative Artificial Intelligence (GenAI) has sparked interest in its application to modeling and simulation (M&S). Large language models (LLMs) in the domain of GenAI have shown remarkable progress in replicating human learning and thinking processes with structured querying. This study explores a framework which leverages LLMs with chain-of-thought (COT) prompting together with network analytics to improve the effectiveness of collaborative group model building. In this context, our research question is: *How can we utilise LLM and network analytics to improve the efficiency and effectiveness of collaborative modelling, without sacrificing the structured stakeholder engagement process?*

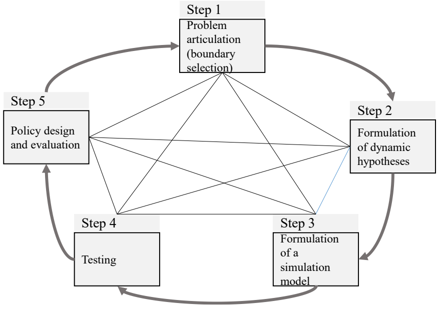
Step 1
Problem articulation (boundary selection)
Step 2
Formulation of dynamic hypotheses
Step 3
Formulation of a simulation model
Step 4
Testing
Step 5
Policy design and evaluation

Figure 1: End-to-end iterative modelling process (Sterman, 2000, p.87);

Methods

Few-shot Chain of Thoughts (COT) prompting can help to direct LLMs to articulate their thought processes before arriving at a final answer[7], [8]. The design of the COT prompting process can leverage on the scaffolding idea, first proposed by Vygotsky [10]. As shown in Table 1, the scaffolded COT prompting process begins with cognitive structuring, where instruction is used to help the LLM understand the tasks and flow of the activity. Cognitive structuring provides explanatory structures that help organize the behaviour of the LLM learner. The next step involves working on a simple case study, where explaining and modeling serve to reduce the degrees of freedom and direct the LLM to appropriate responses. Next, in working with a more complex case study, explaining and modeling serve the dual purposes of ensuring that the LLM stays on target and continues pursuing the intended objective, which is a largely metacognitive function. Finally, to ensure that the LLMs use only correct variable names, the node list is corrected before the edge list are generated for the test case.

Table 1: COT prompting steps with the means and intentions of scaffolding[11]

Step	Description of Step	Intention	Means
1	Understanding the tasks and flow	Cognitive Structuring	Instructing
2	Working on a simple case study (with the provision of list of CLD variables and directed edge lists)	Reduction of Degrees of Freedom	Explaining, Modeling
3	Working on a complex case study (with the provision of list of CLD variables and directed edge lists)	Direction Maintenance, Reduction of Degrees of Freedom	Explaining, Modeling
4	Working on the actual case study (with the provision of list of CLD variables)	Direction Maintenance, Reduction of Degrees of Freedom	Explaining, Modeling

The LLM-augmented process is shown in Figure 2. After the system modelers validate and interpret (Task 2b) the transcripts of stakeholder conversations (produced in Tasks 1 and 2a) , they develop pseudocode (Task 3) which is defined as the simplest possible written textual description of the system. This pseudocode is based on the system modelers' interpretation of the model scoping conversations[15]. The LLMs are then used to first generate system variables (Task 4) in a node list, and then create a list of directed, signed edges representing CLD relationships (Task 5) – "1" for positive links and "-1" for negative links. Modelers then review the output of Task 4, comparing it with their intuition, and then provide their finalized variable list which meets the semantic needs of the context back to the LLMs. This is similar to the student-teacher feedback in scaffolding theory[11] where the human modelers attempt to direct the LLM towards the correct solution for the problem at hand. From this directed graph representation, we use the NetworkX library of Python[16] to produce an exhaustive enumeration of loops (Task 6a) and evaluate the betweenness centrality of nodes (Task 6b). Betweenness centrality plays a role in mediating or brokering role in transmitting effects in the network[17].

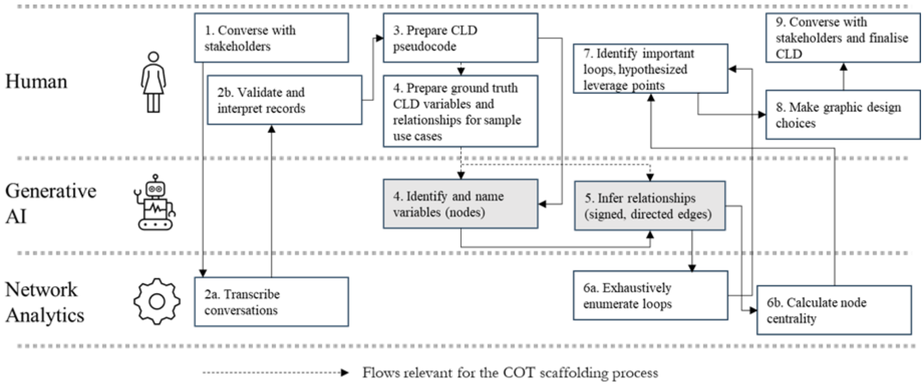
Human
Generative AI
Network Analytics
1. Converse with stakeholders
2b. Validate and interpret records
3. Prepare CLD pseudocode
4. Prepare ground truth CLD variables and relationships for sample use cases
7. Identify important loops, hypothesized leverage points
9. Converse with stakeholders and finalise CLD
8. Make graphic design choices
4. Identify and name variables (nodes)
5. Infer relationships (signed, directed edges)
2a. Transcribe conversations
6a. Exhaustively enumerate loops
6b. Calculate node centrality
Flows relevant for the COT scaffolding process

Figure 2: CLD creation using COT prompting based on scaffolding

We evaluate the hypothesis that such a scaffolded COT approach can achieve sufficiently required accuracy for augmenting human modellers in the modelling process across three LLMs (ChatGPT 4o[12], Claude 3.7 Sonnet[13], and DeepSeek R1[14]) based on precision, recall and F1 scores. The ground truth is the set of CLD variables and relationships developed by experienced system modelers with the inputs of stakeholders.

Findings

Table 2 shows the accuracy of the edge lists generated by the LLMs. Precision, recall and F1 score are the relevant metrics since true negatives have no meaning in this context. The final results showed that ChatGPT 4o[12] and Claude 3.7 Sonnet[13] generated identical edge lists; DeepSeek R1[14] had one edge less than them. For the ChatGPT 4o and Claude 3.7 Sonnet, 12 loops, 6 reinforcing and 6 balancing, were generated. As a result of a missing link, the DeepSeek R1 only has 6 reinforcing loops.

Table 2: Accuracy of Edge List measured against system modelling team

	ChatGPT 4o	Claude 3.7 Sonnet	DeepSeek R1
Precision	1.00	1.00	1.00
Recall	1.00	1.00	0.97
F1 score	1.00	1.00	0.98

Our proposed approach selectively uses LLM technology and combines it with network analytics to produce outputs that can be objectively assessed for accuracy. This approach leaves stakeholders and modelers in charge of condensing a lengthy textual description to a concise one; identifying and naming variables; discussing which of a total set of feedback loops are important in the mental models of those studying the problem; hypothesizing about potential points of leverage (as a precursor to quantitative simulation) and trying out different graphical layouts of the CLD to guide the group model building process with stakeholders.

Conclusion

This study also revealed the potential of widely available LLMs for augmenting some tasks within the model building process. A specific real-world health system case study is used to demonstrate the feasibility of a COT approach based on the scaffolding paradigm in teacher-student interactions. The LLMs utilized in the study can be widely accessible and the network analytics codes are provided for reproducibility of this study[1].

[1] Code and supplementary files can be found: https://github.com/seanlam74/ISDC_GenAI_CLD

References

[1] C. Ebert and P. Louridas, “Generative AI for software practitioners,” *IEEE Softw.*, vol. 40, no. 4, pp. 30–38, 2023.

[2] E. Frydenlund, J. Martínez, J. J. Padilla, K. Palacio, and D. Shuttleworth, “Modeler in a box: how can large language models aid in the simulation modeling process?,” *SIMULATION*, vol. 100, no. 7, pp. 727–749, Jul. 2024, doi: 10.1177/00375497241239360.

[3] N. Ghaffarzadegan, A. Majumdar, R. Williams, and N. Hosseinichimeh, “Generative agent-based modeling: an introduction and tutorial,” *Syst. Dyn. Rev.*, vol. 40, no. 1, p. e1761, Jan. 2024, doi: 10.1002/sdr.1761.

[4] N.-Y. G. Liu and D. Keith, “Leveraging Large Language Models for Automated Causal Loop Diagram Generation: Enhancing System Dynamics Modeling through Curated Prompting Techniques,” *Available SSRN 4906094*, 2024, Accessed: Mar. 05, 2025. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4906094

[5] J. Sterman, *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Boston: Irwin/McGraw-Hill, 2000.

[6] P. S. Hovmand, *Community Based System Dynamics*. New York: Springer, 2014.

[7] B. Wang *et al.*, “Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters,” Jun. 01, 2023, *arXiv*: arXiv:2212.10001. doi: 10.48550/arXiv.2212.10001.

[8] J. Wei *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 24824–24837, 2022.

[9] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” 2020, *arXiv*. doi: 10.48550/ARXIV.2005.14165.

[10] L. S. Vygotsky, *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, 1980. doi: 10.2307/j.ctvjf9vz4.

[11] J. Van De Pol, M. Volman, and J. Beishuizen, “Scaffolding in Teacher–Student Interaction: A Decade of Research,” *Educ. Psychol. Rev.*, vol. 22, no. 3, pp. 271–296, Sep. 2010, doi: 10.1007/s10648-010-9127-6.

[12] OpenAI, *ChatGPT [Large language model]. https://chat.openai.com/chat*. (2023). OpenAI.

[13] Anthropic, *ClaudeAI [Large Language Model]*. (2023). Accessed: Mar. 01, 2025. [Online]. Available: https://www.anthropic.com/

[14] DeepSeek-AI *et al.*, “DeepSeek LLM: Scaling Open-Source Language Models with Longtermism,” 2024, *arXiv*. doi: 10.48550/ARXIV.2401.02954.

[15] R. K. E. Bellamy, “What Does Pseudo-Code Do? A Psychological Analysis of the use of Pseudo-Code by Experienced Programmers,” *Human–Computer Interact.*, vol. 9, no. 2, pp. 225–246, Jun. 1994, doi: 10.1207/s15327051hci0902_3.

[16] A. A. Hagberg, D. A. Schult, and P. J. Swart, “Exploring Network Structure, Dynamics, and Function using NetworkX,” presented at the Python in Science Conference, Pasadena, California, Jun. 2008, pp. 11–15. doi: 10.25080/TCWV9851.

[17] J. McGlashan, M. Johnstone, D. Creighton, K. de la Haye, and S. Allender, “Quantifying a systems map: network analysis of a childhood obesity causal loop diagram,” *PloS One*, vol. 11, no. 10, p. e0165459, 2016.

[18] A. Kumar *et al.*, “Strategizing towards the future hospital: a systems thinking approach,” *Health Res. Policy Syst.*, vol. 23, p. 71, 2025.

[19] U. Brandes, “A faster algorithm for betweenness centrality*,” *J. Math. Sociol.*, vol. 25, no. 2, pp. 163–177, Jun. 2001, doi: 10.1080/0022250X.2001.9990249.

[20] L. Crielaard *et al.*, “Using network analysis to identify leverage points based on causal loop diagrams leads to false inference,” *Sci. Rep.*, vol. 13, no. 1, p. 21046, 2023.

[21] P. Barbrook-Johnson and A. S. Penn, *Systems Mapping: How to build and use causal models of systems*. Springer Nature, 2022. Accessed: Mar. 05, 2025. [Online]. Available: https://library.oapen.org/bitstream/handle/20.500.12657/57376/1/978-3-031-01919-7.pdf

[22] J. Brophy, “Toward a model of the value aspects of motivation in education: Developing appreciation for..,” *Educ. Psychol.*, vol. 34, no. 2, pp. 75–85, Mar. 1999, doi: 10.1207/s15326985ep3402_1.

[23] S. Hao *et al.*, “Reasoning with Language Model is Planning with World Model,” Oct. 23, 2023, *arXiv*: arXiv:2305.14992. doi: 10.48550/arXiv.2305.14992.

[24] P. S. Hovmand, *Community Based System Dynamics*. New York, NY: Springer New York, 2014. doi: 10.1007/978-1-4614-8763-0.

[25] T. Xie, T. Yin, V. Keshava, X. Zhang, and S. R. Jonnalagadda, “BiasCause: Evaluate Socially Biased Causal Reasoning of Large Language Models,” 2025, *arXiv*. doi: 10.48550/ARXIV.2504.07997.

[26] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada: ACM, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[27] J.-F. Ton, M. F. Taufiq, and Y. Liu, “Understanding Chain-of-Thought in LLMs through Information Theory,” in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=IjOWms0hrf

[28] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang, “An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning,” 2023, *arXiv*. doi: 10.48550/ARXIV.2308.08747.

[29] L. Chen, M. Zaharia, and J. Zou, “How is ChatGPT’s behavior changing over time?,” 2023, *arXiv*. doi: 10.48550/ARXIV.2307.09009.