Semiquantitative in-context causal reasoning for LLMs

Marcelo R. Thielo[1], **Bernardo N. Rocha** [1]

[1]Campus Alegrete – Universidade Federal do Pampa (Unipampa)
Av. Tiaraju, 810 - Ibirapuitã, Alegrete - RS, 97546-550

{marcelothielo, bernardorocha.aluno}@unipampa.edu.br

Abstract. *We propose a novel algorithm that employs simulations of causal dynamical models and fuzzy logic to push forward the ability of large language models(LLMs) on extracting networks of causal relations from text documents. As a means to that, we explored the ability of LLMs to emulate actual critical reasoning and self-criticize on their own conclusions based on simulations of the gradually improving causal models, which we called In-Context Reasoning. As the self-assessment strategy, an instance of LLM-as-a-judge was employed. The results are encouraging, and we expect to further contribute to the investigations of the potential of LLMs in activities that require more accurate logical and structured reasoning.*

1. Introduction

In recent years, large language models (LLMs) have gained prominence in both academic research and commercial applications, driven by their ability to generate and interpret natural language text with unprecedented levels of fluency and contextual awareness. However, despite these advances, such architectures exhibit substantial limitations when applied to tasks requiring causal reasoning, logical inference, or the generation of reliable knowledge. One of the main challenges lies in the absence of grounding mechanisms[Liu 2023], that is, the ability to anchor LLM responses to trustworthy models of the real world. This often leads to hallucinations [Xu et al. 2024], or erroneous outputs caused by insufficient domain-specific training, dataset bias, or architectural nondeterminism.

To address this issue, the present work proposes the approach of **Semiquantitative In-Context Reasoning**, an extension of the In-Context Learning (ICL) technique [Dong et al. 2024]. In our approach, the LLM is engaged in an iterative cycle of generation, simulation, interpretation, and refinement of semiquantitative causal models. Within this framework, the goal is to assess the feasibility of using LLMs as active agents in knowledge discovery, combining natural language, semi-quantitative model simulation, and algorithmic verification. As a distinguishing feature, the proposed method incorporates self-evaluation cycles, inspired by approaches such as CRITIC [Gou et al. 2024] and LLM-as-a-judge [Li et al. 2025], allowing the causal model to evolve based on the LLM's own analysis of its outputs. This enables an evaluation of the approach, its differences from conventional methods, and whether the conceptual models anchor the LLM to more coherent and grounded responses.

2. Proposed Approach

Our approach, called **In-Context Reasoning** (ICR), extends In-Context Learning (ICL) by integrating causal models generation with numerical simulations and interpretation-based feedback. Its novelty lies in combining (1) a two-LLM architecture (a generator and a judge), and (2) an embedded simulation-based interpretation as a self-correction technique. The summarized workflow can be seen in Figure 1, which has its steps described below:

- Step 1: **Model Generation**: The primary LLM extracts variables and causal links from the input text, producing a digraph for the initial causal model, with weights and initial values translated into a system of ordinary differential equations (ODEs) in the form $\dot{x_i} = f(\mathbf{W}^T\mathbf{X})$, where $\mathbf{W}$ and $\mathbf{X}$ are vectors containing weights and values as a neural network, and i ranges from 1 to the total number of variables from the text, as identified by the LLM.
- Step 2 : **Simulation**: The ODEs are solved numerically, simulating the current model dynamics over time.
- Step 3: **Interpretation**: The behavior of the system is summarized into a textual description through fuzzy variables and provided to the feedback stage (e.g. "The variable A is low and growing fast while the variable B is high and decreasing slowly.")
- Step 4: **Self-Critique**: The judge model uses the base context and the interpretation to generate concise feedback on the coherence and plausibility of the model and its simulation.
- Step 5: **Refinement**: The primary model revises the model according to the feedback.

The cycle repeats until reaching either the iterations limit or no changes between two successive models.

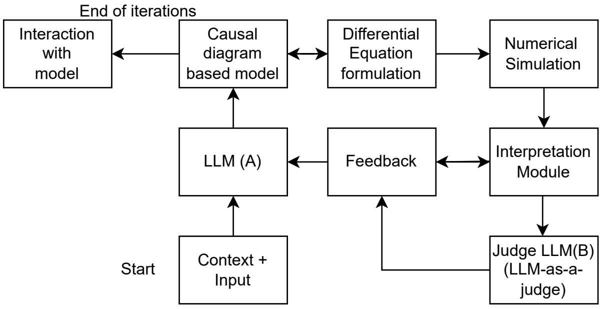


Figure 1. The proposed architecture for semiquantitative in-context reasoning.

For the implementation of the experiments, the Gemini API (*genai* library in Python) was utilized, employing different models and instances for different roles. The primary LLM for generating models used is Gemini 2.5 Pro, because of the properties of Large Reasoning Models and the precision and coherence observed in our preliminary experiments. The judge model (*LLM-as-a-judge*) used is Gemini 2.5 Flash, due to the less demanding function of judging the models.

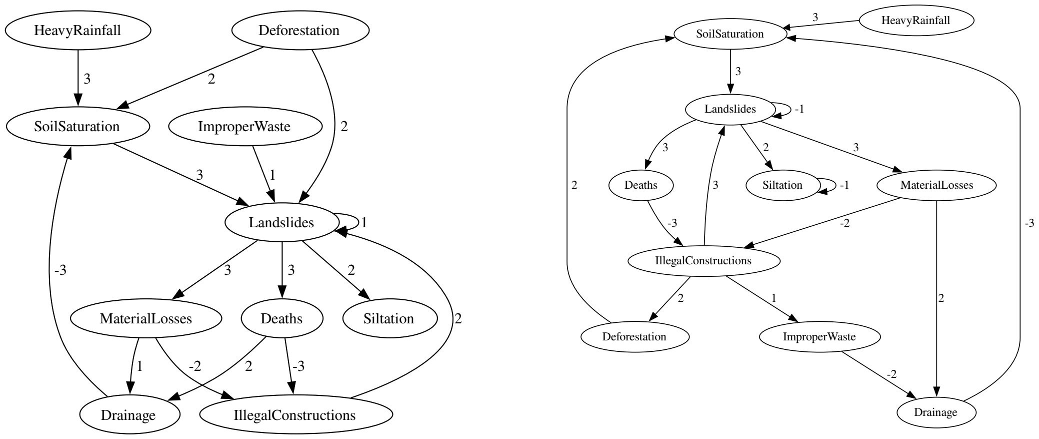
3. Experiments

The experimental phase involved 18 preliminary and 22 structured tests typically running in 3 iterations. This limitation was primarily due to the context window constraints of the LLMs employed and the design choice of maintaining a persistent instance for the generator model throughout the iterative process. For the generator model, our experiments involved crafting instructions to produce digraphs of causal models represented with the graphviz language [Gansner and North 2000]. Principles like Occam's Razor ("Prefer simpler models with the same meaning") were integrated into our prompt design, as the tests showed simplification and more concise variables and relationships.

For the judge model, prompts included the base context, the algorithmic interpretation of the simulation results, and concise behavioral instructions (e.g., "Be succinct", "Limit your output to feedback on the generated model"). This structure was designed to elicit focused feedback on the model's coherence and behavior.

4. Results and Discussion

Some visible improvements across iterations indicate the approach's potential for producing increasingly coherent models. As seen in the diagrams of Figure 2, in the experiment with a text about landslides [Brasil Escola 2024], variables like "ImproperWaste" and "Deforestation" were initially created without proper connections in the model, a structural inconsistency corrected in later iterations through the feedback loop. We observed that incorporating the simulation-based interpretation into the LLM-as-a-judge prompt further enhanced consistency and refinements. Average digraph density for the models with interpretation feedback was slightly higher ($\bar{D} = 0.163$) than for models without it ($\bar{D} = 0.156$). Standard deviations were $\sigma = 0.052$ and $\sigma = 0.049$ respectively, suggesting coherence in the results.



(a) First model generated **(b)** Model in second iteration

Figure 2. Causal diagrams for models obtained in two subsequent iterations

The Figure 3 shows how the interpretation step made the models dynamically more consistent through the iterations as seen by the comparison of the simulation outputs.

An evaluation of the feasibility and effectiveness of In-Context Reasoning was conducted through a series of structured experiments, aiming to guide LLMs toward more coherent and consistent outputs. Metrics such as density and others were collected, but part of the evaluation is still done by direct model inspection so as to guarantee no mistakes from the LLM.

5. Conclusion and Future work

The results have been encouraging across diverse test scenarios, highlighting the potential of structured feedback loops for enhancing the quality and coherence of causal models generated by LLMs. The main findings at this point are summarized below:

- The iterative simulation-interpretation loop seems to have automatically improved diagram coherence across different inputs and prompt structures and also identified and corrected internal model inconsistencies.
- Some degree of misinterpretation of delayed causal effects remains, and LLM architectural constraints such as context window size and non-determinism still affect consistency and scalability.

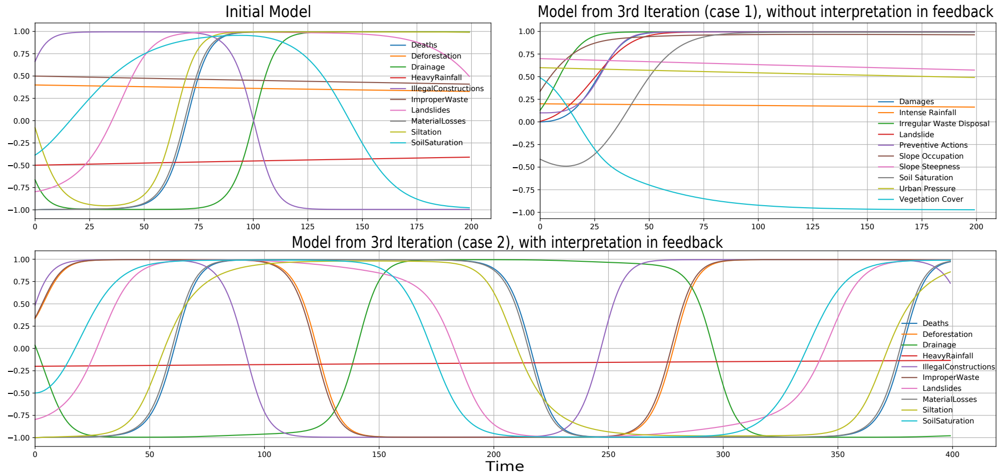


Figure 3. Dynamical behavior of the models' variables. Y axis is semiquantitative: values ≈ 1 mean high, ≈ -1 mean low. Derivatives > 0 mean increasing, < 0 decreasing or ≈ 0 stability.

Future work will focus on extending the iteration depth, experimenting with domain-specific judges, refining prompt strategies, and incorporating the corrected causal knowledge into queries' responses about the text input. Also planned are tests with larger datasets and collecting more sophisticated metrics such as structural coherence, graph complexity, and causal precision.

References

Brasil Escola (2024). Deslizamentos de encostas. `https://brasilescola.uol.com.br/geografia/deslizamentos-encostas.htm`. Brasil Escola (UOL Educação). Accessed: 2025-08-05.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Liu, T., Chang, B., Sun, X., Li, L., and Sui, Z. (2024). A survey on in-context learning.

Gansner, E. R. and North, S. C. (2000). An open graph visualization system and its applications to software engineering. *Software: Practice and Experience*, 30(11):1203–1233.

Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., and Chen, W. (2024). Critic: Large language models can self-correct with tool-interactive critiquing.

Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., Shu, K., Cheng, L., and Liu, H. (2025). From generation to judgment: Opportunities and challenges of llm-as-a-judge.

Liu, B. (2023). Grounding for artificial intelligence.

Xu, Z., Jain, S., and Kankanhalli, M. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. *arXiv preprint arXiv:2401.11817*.