Network Dynamics Reasoning: A Novel Benchmark for Evaluating Multi-Step Inference in Large Language Models

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

We introduce a novel benchmark for evaluating large language models' ability to reason about network dynamics and multi-step system evolution. Our benchmark tests models on predicting the final state of threshold-based adoption processes in social networks, requiring precise numerical prediction after complex temporal reasoning. We evaluate five state-of-the-art models across different architectures and API providers, revealing significant performance gaps and emergent reasoning capabilities. Our key findings show that Google's Gemini models substantially outperform Meta's Llama and Google's Gemma models, with Gemini 1.5 Pro achieving 55% accuracy compared to 10% for Llama 3.3 70B, despite the latter's larger parameter count. This benchmark addresses critical gaps in current LLM evaluation by testing contamination-resistant synthetic scenarios, precise numerical reasoning, and multi-step temporal dynamics—capabilities essential for AI systems operating in complex real-world environments.

1 Introduction

Current large language model (LLM) evaluation benchmarks primarily focus on static knowledge retrieval, reading comprehension, and single-step reasoning tasks. However, real-world applications increasingly require AI systems to understand and predict the evolution of complex dynamic systems over multiple time steps. Network dynamics—the study of how behaviors, information, or states propagate through interconnected systems—represents a fundamental reasoning challenge that remains underexplored in LLM evaluation literature.

We present a novel benchmark that evaluates LLMs' ability to predict the final states of threshold-based adoption processes in social networks. Our approach addresses several critical limitations in current evaluation methodologies: (1) **Contamination resistance**: synthetic, deterministic scenarios eliminate training data contamination concerns; (2) **Precise numerical reasoning**: requires exact integer predictions rather than multiple-choice or qualitative responses; (3) **Multi-step temporal reasoning**: models must trace system evolution across multiple discrete time steps; and (4) **Emergent behavior prediction**: understanding when and why cascade effects occur in networked systems.

The benchmark tests models on 60 deterministic threshold-adoption scenarios across varying network topologies, threshold patterns, and initial conditions. Each scenario requires models to predict the exact final number of adopters after the system reaches equilibrium. Our evaluation reveals dramatic performance differences between model families, with implications for understanding emergent reasoning capabilities and scaling laws in contemporary LLMs.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2 Related Work

LLM Evaluation and Benchmarking. Existing benchmarks like MMLU [5], BIG-bench [8], and HELM [6] primarily evaluate static knowledge and single-step reasoning. Recent work has begun exploring multi-step reasoning through mathematical problem-solving [3] and scientific reasoning [7], but temporal dynamics remain understudied.

Network Dynamics and Threshold Models. Threshold models of social influence, pioneered by Granovetter [4] and formalized by Watts [9], provide a rich framework for studying collective behavior. These models exhibit complex emergent phenomena including cascade effects and critical thresholds, making them ideal for testing AI reasoning capabilities.

Synthetic Evaluation and Data Contamination. Growing concerns about training data contamination [1, 2] motivate synthetic evaluation approaches. Our deterministic scenario generation ensures contamination resistance while maintaining reproducibility.

3 Methodology

3.1 Threshold-Adoption Model

We implement a deterministic threshold-adoption process on fixed network topologies. Each agent i has a binary state $s_i(t) \in \{0, 1\}$ (non-adopter/adopter) and threshold θ_i. The update rule is:

$$s_i(t+1) = \begin{cases} 1 & \text{if } s_i(t) = 1 \text{ or } \sum_{j \in N(i)} s_j(t) \geq \theta_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $N(i)$ denotes agent i's neighbors. Updates are synchronous and adoption is irreversible (monotonic dynamics).

3.2 Scenario Generation

We generate 60 deterministic scenarios through Cartesian product combinations:

- **Network topologies**: Ring networks with 0, 1, 2, or 3 chord shortcuts
- **Threshold patterns**: Constant, cyclic, split, and alternating assignment strategies
- **Initial conditions**: 1-4 seed adopters in contiguous or spaced arrangements

Each scenario involves 12 agents and runs for up to 12 time steps (convergence typically occurs within 2-6 steps). This design ensures scenario diversity while maintaining tractable complexity for detailed model analysis.

3.3 Evaluation Protocol

Models receive self-contained prompts including: (1) explicit dynamics rules, (2) complete network topology, (3) individual agent thresholds, (4) initial adopter configuration, and (5) first two time steps of system evolution. The task requires predicting the final adopter count after convergence.

We employ exact integer matching for scoring—no partial credit is awarded. This strict criterion ensures evaluation precision and distinguishes between models that truly understand the dynamics versus those making educated guesses.

3.4 Model Selection and API Configuration

We evaluate five contemporary models across different architectures and API providers:

- **Google Gemini**: 1.5 Flash, 1.5 Pro (via Google AI API)
- **Meta Llama**: 3.3 70B Versatile, 3.1 8B Instant (via Groq API)
- **Google Gemma**: 2 9B IT (via Groq API)

This selection enables comparison across model families, parameter scales, and API optimization strategies.

4 Results

4.1 Overall Performance

Table 1 shows dramatic performance variation across models. Gemini 1.5 Pro achieves the highest accuracy at 55%, followed by Gemini 1.5 Flash at 25%. Open models perform significantly worse, with Llama 3.3 70B at 10%, Gemma2 9B at 5%, and Llama 3.1 8B showing particularly poor performance at 0%.

Model	Provider	Parameters	Accuracy	Correct/Total
Gemini 1.5 Pro	Google	-	55.0%	33/60
Gemini 1.5 Flash	Google	-	25.0%	15/60
Llama 3.3 70B	Groq	70B	10.0%	6/60
Gemma2 9B	Groq	9B	5.0%	3/60
Llama 3.1 8B	Groq	8B	0.0%	0/60

Table 1: Model performance on network dynamics benchmark.

4.2 Architecture vs. Scale Analysis

A striking finding emerges when comparing model performance against parameter count. Gemini 1.5 Flash significantly outperforms Llama 3.3 70B (25% vs. 10%) despite likely having fewer parameters. This suggests that **architectural design and training methodology matter more than raw scale** for complex reasoning tasks involving temporal dynamics.

4.3 API Provider Effects

Models accessed via different API providers show systematic performance differences. Google's direct API consistently delivers superior performance compared to Groq-accelerated variants, even when comparing identical model architectures (Gemini vs. Gemma). This raises questions about whether optimization for inference speed may compromise reasoning capabilities.

4.4 Error Pattern Analysis

Qualitative analysis reveals distinct failure modes:

- **Gemini models**: Conservative predictions within realistic bounds, occasional off-by-one errors
- **Llama 3.3 70B**: Frequent extreme predictions (0 or 12), suggesting binary thinking
- **Llama 3.1 8B**: Wild predictions (91, 92, 100) indicating poor task comprehension
- **Gemma2 9B**: Variable prediction quality with some correct reasoning

5 Discussion

5.1 Implications for Emergent Abilities

Our results provide evidence for emergent reasoning capabilities that appear discontinuously across model families rather than smoothly with scale. The dramatic performance gap between Gemini Pro (55%) and all other models suggests qualitative differences in network reasoning abilities that cannot be explained by parameter count alone.

5.2 Scaling Laws and Architecture

The superior performance of smaller Gemini models over larger Llama models challenges simple parameter-count scaling assumptions. This aligns with recent findings that specialized capabilities may require specific architectural innovations rather than merely increasing model size.

5.3 Benchmark Validity and Future Directions

The benchmark's difficulty level (best performance: 55%) indicates substantial room for improvement, avoiding ceiling effects common in saturated benchmarks. The contamination-resistant design and precise numerical evaluation provide robust assessment of genuine reasoning capabilities.

Future work could explore: (1) prompt engineering strategies to improve performance, (2) few-shot learning with worked examples, (3) hybrid approaches combining neural and symbolic methods, and (4) scaling to larger networks and more complex dynamics.

5.4 Limitations

Our evaluation has several limitations: (1) relatively small scenario set (60 cases), (2) focus on one specific network dynamic, (3) binary adoption states, and (4) fixed network size (12 agents). Future work could address these limitations by expanding to larger scenario sets, multiple dynamic types, and variable network sizes.

6 Conclusion

We introduced a novel benchmark for evaluating LLMs' ability to reason about network dynamics and multi-step temporal processes. Our evaluation of five state-of-the-art models reveals significant performance gaps that correlate more strongly with model family and architecture than with parameter count. These findings have important implications for understanding emergent reasoning capabilities, scaling laws, and the development of AI systems capable of operating in complex dynamic environments.

The benchmark addresses critical gaps in current LLM evaluation methodology through contamination-resistant synthetic scenarios, precise numerical assessment, and multi-step reasoning requirements. We hope this work contributes to more comprehensive and robust evaluation protocols for the evolving landscape of large language models.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Sastry Girish, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[3] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[4] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

[5] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[6] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[7] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[8] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[9] Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the national academy of sciences*, 99(9):5766–5771, 2002.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our contribution of a novel network dynamics benchmark and accurately represent our findings about performance gaps between model families.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed and suggest directions for future research?

   Answer: [Yes]

   Justification: Section 4.4 explicitly discusses limitations including scenario set size, focus on single dynamic type, binary states, and fixed network size.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This is an empirical evaluation paper without theoretical contributions requiring formal proofs.

4. **Experimental Result Reproducibility**

   Question: Does the paper provide sufficient details to reproduce the experimental results?

   Answer: [Yes]

   Justification: We provide complete methodology including formal model definition, scenario generation procedure, evaluation protocol, and model specifications. The deterministic nature ensures full reproducibility.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code needed to reproduce the experimental results?

   Answer: [Yes]

   Justification: Code and benchmark data will be made available upon publication to ensure reproducibility and enable future research.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, number of runs, etc)?

   Answer: [Yes]

   Justification: We provide complete experimental details including scenario generation, evaluation protocol, model configurations, and API specifications.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and explain how they are derived (e.g., independent runs, std over k-folds, etc)?

   Answer: [No]

   Justification: Given the deterministic nature of our benchmark and the focus on architectural comparisons rather than statistical inference, error bars are not applicable. Each model receives identical evaluation scenarios.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information about the computational resources used?

Answer: [Yes]

Justification: We specify API providers, model access methods, and evaluation scope (180 total API calls across 5 models).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research involves only synthetic benchmark evaluation of publicly available models without any ethical concerns.

10. **Broader Impacts**

    Question: Does the paper discuss any potential broader impact of the work?

    Answer: [Yes]

    Justification: The discussion section addresses implications for AI development, network science applications, and practitioner guidance.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been taken to ensure that the work is responsible and safe?

    Answer: [Yes]

    Justification: Our evaluation uses only synthetic scenarios and publicly available models, eliminating potential safety concerns from real-world data or novel model development.