

Reasoning Topology Matters: A Controlled Study of LLM-Based Cybersecurity Analysis

1st Jiling Zhou^{#,*}

Department of Computing
University of Turku
Turku, Finland

ORCID: 0009-0006-7903-9355

*Corresponding author.

2nd Aisvarya Adeseye[#]

Department of Computing
University of Turku
Turku, Finland

ORCID: 0009-0003-2401-3076

[#]These authors contributed equally.

3rd Antti Hakkala

Department of Computing
University of Turku
Turku, Finland

ORCID: 0000-0002-0932-7814

4th Seppo Virtanen

Department of Computing
University of Turku
Turku, Finland

ORCID: 0000-0002-9487-3018

5th Jouni Isoaho

Department of Computing
University of Turku
Turku, Finland

ORCID: 0000-0002-5789-3992

Abstract—Large Language Models (LLMs) are increasingly used in cybersecurity, where accurate analysis often requires multi-step and context-dependent reasoning over complex and heterogeneous data. However, existing prompting approaches typically focus on eliciting reasoning without explicitly considering how intermediate reasoning steps are structurally organized. We introduce *Security Reasoning Topology*, which models reasoning through three representative structures: Linear, Branching, and Graph. To evaluate their effects, we conduct controlled experiments on three cybersecurity datasets covering MITRE ATT&CK network traffic, cyber threat intelligence (CTI), and CVE vulnerability analysis. We evaluate multiple LLMs, including Llama 2 (7B, 13B, 70B), GPT-5.1, and Mistral Large 3, while keeping task inputs consistent and controlling reasoning structure through system-level prompting. Results show that reasoning topology substantially affects performance: Graph reasoning achieves the highest overall accuracy, improving over few-shot prompting by 9.8–12.2 percentage points across datasets, while Branching provides a strong intermediate solution. The results further show that the effect of reasoning topology remains consistent across model families and scales, highlighting reasoning topology as an important design factor for LLM-based cybersecurity analysis.

Index Terms—Large Language Models, Cybersecurity, Reasoning Topology, Structured Reasoning

I. INTRODUCTION

Large Language Models (LLMs) are increasingly used in cybersecurity tasks such as threat detection, vulnerability analysis, cyber threat intelligence (CTI), and incident response [1]–[4]. In these settings, the quality of model reasoning can directly affect how security evidence is interpreted and how analytical conclusions are formed [5]–[7]. Cybersecurity tasks, however, differ substantially in their reasoning requirements. Vulnerability analysis often follows sequential cause–effect relationships, CTI analysis may require considering multiple competing explanations, while attack-path analysis requires reasoning over interconnected behaviors and dependencies. Applying the same reasoning strategy across these structurally different tasks may therefore limit analytical performance.

Existing work has extensively studied prompting and reasoning methods such as Chain-of-Thought (CoT), Tree-of-Thought (ToT), and Graph-of-Thought (GoT) [8]–[11], while cybersecurity research has largely evaluated LLMs on individual security applications. However, these two directions are typically studied separately: reasoning methods are treated as independent prompting techniques, and limited attention has been given to the *structural organization* of reasoning across cybersecurity tasks. This leaves an important question unresolved: how does the structure of intermediate reasoning itself influence cybersecurity analysis?

To investigate this question, we introduce *Security Reasoning Topology*, a unified structural view of LLM reasoning. We characterize three representative topologies: **Linear**, which follows a sequential reasoning path; **Branching**, which explores multiple candidate paths; and **Graph**, which supports interconnected dependencies, information reuse, and iterative refinement. These structures are instantiated using CoT, ToT, and GoT, respectively. We evaluate them under a controlled setting across three cybersecurity datasets covering MITRE ATT&CK network traffic, CTI, and CVE vulnerability analysis. To isolate the effect of reasoning topology, task-specific user inputs are kept consistent while system-level instructions control the reasoning structure across models.

Our study addresses two research questions: *RQ1: How does reasoning topology affect LLM performance across cybersecurity analysis tasks?* and *RQ2: How consistent are topology effects across model scales, and how does prompt organization affect performance under topology-based reasoning?* Experiments across multiple LLM families and model scales show that reasoning topology has a measurable impact on performance. Graph reasoning achieves the strongest overall accuracy, while Branching provides an intermediate solution between simpler Linear reasoning and more complex Graph reasoning.

path, where each step primarily depends on its predecessor. **Branching reasoning** generates and evaluates multiple candidate paths, enabling comparison among alternative hypotheses. **Graph reasoning** allows states to be interconnected, reused, merged, and refined, supporting reasoning over more complex dependencies.

These topologies provide a common structural abstraction for representative LLM reasoning paradigms. In our implementation, Linear, Branching, and Graph reasoning are instantiated using Chain-of-Thought (CoT), Tree-of-Thought (ToT), and Graph-of-Thought (GoT), respectively [8]–[10]. Rather than assuming a fixed mapping between cybersecurity tasks and optimal topologies, we empirically evaluate how reasoning-structure complexity affects performance across heterogeneous security tasks and whether these effects remain consistent across model scales.

IV. EXPERIMENTAL SETUP

A. Datasets

We evaluate reasoning topologies on three publicly available cybersecurity datasets covering heterogeneous data and task structures. The **ATT&CK dataset** contains structured network traffic records annotated with MITRE ATT&CK tactic labels, including protocol, service, connection state, packet and byte counts, ports, and timestamps [19]. The **CTI dataset** contains 1,100 threat reports with threat categories and indicators of compromise, providing unstructured textual inputs for threat analysis [20]. The **CVE dataset** contains structured vulnerability records with identifiers, categories, and technical descriptions [21]. Duplicate records are removed and samples are balanced where necessary to reduce dataset bias.

B. Compared Methods

We compare three topology-based reasoning methods with three non-topological baselines. The topology-based methods instantiate **Linear**, **Branching**, and **Graph** reasoning using CoT, ToT, and GoT, respectively. The baselines capture three complementary factors unrelated to explicit reasoning topology: **few-shot prompting** provides labeled demonstrations before prediction [22]; **structured JSON prompting** constrains the output to a predefined schema [23]; and **RAG** augments the model input with retrieved cybersecurity context such as ATT&CK descriptions, CTI evidence, or CVE information [24]. This design allows us to compare explicit reasoning structure with improvements arising from demonstrations, output constraints, and external knowledge.

C. Controlled Prompt Design and Models

To control topology implementation, we separate reasoning instructions from task-specific inputs. The **system prompt** specifies how intermediate reasoning states should be organized for Linear, Branching, or Graph reasoning, while the **user prompt** contains the task instruction and security input. Within the topology-based conditions, the task formulation is kept fixed while only the system-level reasoning structure

TABLE I
AGGREGATE ACCURACY ACROSS FIVE EVALUATED MODELS AND FIVE INDEPENDENT RUNS ON THE THREE CYBERSECURITY DATASETS. VALUES ARE REPORTED AS MEAN $\pm$ STANDARD DEVIATION.

Method	ATT&CK	CTI	CVE
Few-shot	75.2 $\pm$ 1.8	72.8 $\pm$ 2.1	70.2 $\pm$ 2.0
JSON	77.2 $\pm$ 1.7	76.2 $\pm$ 1.9	73.4 $\pm$ 1.8
RAG	80.6 $\pm$ 1.6	80.8 $\pm$ 1.7	76.8 $\pm$ 1.7
CoT	78.2 $\pm$ 1.6	78.2 $\pm$ 1.8	75.2 $\pm$ 1.7
ToT	82.2 $\pm$ 1.5	82.2 $\pm$ 1.6	79.2 $\pm$ 1.6
GoT	85.0 $\pm$ 1.4	85.0 $\pm$ 1.5	82.0 $\pm$ 1.5

changes, allowing the effect of topology to be compared under consistent inputs.

The topology implementations are controlled at the prompt level rather than through a separate external orchestration framework; the system prompt specifies the structural instructions appropriate to each reasoning topology, including branching, dependency, and refinement when applicable.

We evaluate Llama 2 models with 7, 13, and 70 billion parameters, GPT-5.1, and Mistral Large 3 to examine whether topology effects persist across model families and scales. Temperature and top- p are fixed at 0.2 and 0.9, respectively. The maximum generation length is 1024 tokens, extended to 2048 for ToT and GoT to accommodate expanded reasoning structures. Temperature 0.2 was selected through a pilot ablation over $\{0.0, 0.2, 0.5\}$. For each task instance, model outputs are recorded under the same evaluation protocol.

D. Evaluation Metrics

We use **Accuracy** as the primary evaluation metric, measuring the proportion of predictions that match the ground-truth task label. Each model–method–dataset configuration is evaluated over five independent runs. Aggregate results are reported across the five evaluated models and five runs. Paired bootstrap tests are used to evaluate the statistical significance of performance differences between methods.

V. RESULTS AND DISCUSSION

A. Overall Performance

Table I summarizes aggregate accuracy across the five evaluated models and five independent runs on the three cybersecurity datasets. All topology-based methods improve over direct few-shot prompting, while the magnitude of improvement generally increases with reasoning-structure complexity. GoT achieves the highest mean accuracy on all three datasets, reaching 85.0% on ATT&CK and CTI and 82.0% on CVE. Relative to few-shot prompting, these correspond to improvements of 9.8, 12.2, and 11.8 percentage points, respectively.

RAG is the strongest non-topological baseline, achieving 80.6%, 80.8%, and 76.8% accuracy on ATT&CK, CTI, and CVE, respectively. Nevertheless, GoT remains consistently stronger, indicating that explicitly organizing intermediate reasoning can provide benefits beyond supplying additional

TABLE II
AVERAGE ACCURACY ACROSS THE THREE DATASETS FOR DIFFERENT
MODELS AND REASONING TOPOLOGIES.

Model	CoT	ToT	GoT
Llama 2 7B	65.7	69.7	72.7
Llama 2 13B	71.3	75.3	78.3
Llama 2 70B	79.3	83.3	86.3
GPT-5.1	86.3	90.3	93.3
Mistral Large 3	83.3	87.3	89.3

external knowledge. ToT provides a strong intermediate topology, while CoT produces smaller but consistent gains over few-shot prompting.

Paired bootstrap tests further support these differences. GoT significantly outperforms few-shot prompting on all three datasets ($p < 0.001$) and also improves over RAG ($p < 0.05$).

B. Model-Scale Effects

Table II shows that GoT consistently achieves the highest average accuracy within every evaluated model group, followed by ToT and CoT. Within the Llama 2 family, absolute performance increases from 7B to 70B while the ordering of the three topologies remains stable. The same ordering is observed for GPT-5.1 and Mistral Large 3, indicating that the topology effect is consistent across different model families and scales.

C. Prompt Organization Ablation

We further examine whether system–user prompt separation affects the performance of topology-based reasoning. For each topology, we compare three settings: using only a user prompt, merging topology and task instructions into a single user prompt, and separating topology control into the system prompt while keeping task-specific inputs in the user prompt.

The separated design consistently achieves the highest accuracy. Averaged across the three datasets, separating system and user prompts improves accuracy over the user-only setting by 1.9 percentage points for CoT, 2.9 points for ToT, and 3.7 points for GoT. The larger gain for GoT suggests that more complex reasoning structures particularly benefit from explicit system-level topology control. These results support system–user prompt decomposition as a controlled implementation mechanism for reasoning topology.

D. Discussion

Overall, the results show that reasoning topology is a consistent factor in LLM-based cybersecurity analysis. Linear reasoning provides a simple structured approach, Branching improves performance through alternative-path exploration, and Graph reasoning achieves the strongest overall results across the evaluated tasks and models. At the same time, the prompt ablation shows that reliable topology implementation depends on how structural instructions are separated from task-specific inputs. Together, these findings support reasoning topology as a useful abstraction for systematically studying and controlling LLM reasoning in cybersecurity.

Two limitations are particularly relevant. First, the current evaluation focuses on analytical performance and does not explicitly quantify token consumption or end-to-end inference latency; a fuller analysis of accuracy–cost trade-offs is left for future work. Second, we do not systematically categorize topology-specific failure modes. Possible failure modes include ineffective branch selection in Branching reasoning and inconsistent or hallucinated dependencies in Graph reasoning, which warrant further investigation.

VI. CONCLUSION

This work investigates reasoning topology as a structural factor in LLM-based cybersecurity analysis. We formulate Linear, Branching, and Graph reasoning as three representative topologies and instantiate them using CoT, ToT, and GoT under a controlled prompt design. Experiments across three cybersecurity datasets and multiple LLMs show that reasoning topology consistently affects analytical performance, with Graph reasoning achieving the highest overall accuracy and Branching providing a strong intermediate approach. The results also show that these effects persist across model scales and that separating system-level topology instructions from task-specific user inputs improves performance under topology-based reasoning. Overall, the findings suggest that reasoning structure should be treated as an explicit design dimension when developing LLM-based cybersecurity analysis systems. Future work will investigate adaptive topology selection and extend the evaluation to broader security scenarios.

ACKNOWLEDGMENT

This project has received funding from the European Union’s Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101177564 — HAIF.

REFERENCES

- [1] J. Zhang, H. Bu, H. Wen, Y. Liu, H. Fei, R. Xi, L. Li, Y. Yang, H. Zhu, and D. Meng, “When LLMs meet cybersecurity: A systematic literature review,” *Cybersecurity*, vol. 8, no. 1, p. 55, Feb. 2025.
- [2] M. A. Ferrag, F. Alwahedi, A. Battah, B. Cherif, A. Mechri, and N. Tihanyi, “Generative AI and Large Language Models for Cyber Security: All Insights You Need,” Rochester, NY, Jun. 2024.
- [3] H. Xu, S. Wang, N. Li, K. Wang, Y. Zhao, K. Chen, T. Yu, Y. Liu, and H. Wang, “Large Language Models for Cyber Security: A Systematic Literature Review,” *ACM Trans. Softw. Eng. Methodol.*, Sep. 2025.
- [4] H. Jelodar, S. Bai, P. Hamed, H. Mohammadian, R. Razavi-Far, and A. Ghorbani, “Large Language Model (LLM) for Software Security: Code Analysis, Malware Analysis, Reverse Engineering,” Apr. 2025.
- [5] J. Jin, B. Tang, M. Ma, X. Liu, Y. Wang, Q. Lai, J. Yang, and C. Zhou, “Crimson: Empowering Strategic Reasoning in Cybersecurity through Large Language Models,” in *2024 5th International Conference on Computer, Big Data and Artificial Intelligence (ICCBDAI)*, Nov. 2024, pp. 18–24.
- [6] H. F. Atlam, “LLMs in Cyber Security: Bridging Practice and Education,” *Big Data and Cognitive Computing*, vol. 9, no. 7, Jul. 2025.
- [7] Q. He, C. Zhu, S. A. Abahusseini, M. Wang, and M. Zhu, “CoTSentry: Advanced Network Attack Detection with Chain-of-Thought Reasoning,” in *Knowledge Science, Engineering and Management*, T. Zhu, W. Zhou, and C. Zhu, Eds. Singapore: Springer Nature, 2026, pp. 47–58.

- [8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 24 824–24 837.
- [9] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of Thoughts: Deliberate Problem Solving with Large Language Models," in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 11 809–11 822.
- [10] M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nyczyk, and T. Hoefler, "Graph of Thoughts: Solving Elaborate Problems with Large Language Models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, pp. 17 682–17 690, Mar. 2024.
- [11] Y. Yao, Z. Li, and H. Zhao, "GoT: Effective Graph-of-Thought Reasoning in Language Models," in *Findings of the Association for Computational Linguistics: NAACL 2024*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 2901–2921.
- [12] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, and H. Chen, "Reasoning with Language Model Prompting: A Survey," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 5368–5393.
- [13] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-Consistency Improves Chain of Thought Reasoning in Language Models," Mar. 2023.
- [14] Z. Bi, K. Han, C. Liu, Y. Tang, and Y. Wang, "Forest-of-Thought: Scaling Test-Time Compute for Enhancing LLM Reasoning," Apr. 2025.
- [15] A. Karras, L. Theodorakopoulos, C. Karras, A. Theodoropoulou, I. Kallianpakou, and G. Kalogeratos, "LLMs for Cybersecurity in the Big Data Era: A Comprehensive Review of Applications, Challenges, and Future Directions," *Information*, vol. 16, no. 11, Nov. 2025.
- [16] F. Perrina, F. Marchiori, M. Conti, and N. V. Verde, "AGIR: Automating Cyber Threat Intelligence Reporting with Natural Language Generation," in *2023 IEEE International Conference on Big Data (BigData)*, Dec. 2023, pp. 3053–3062.
- [17] M. Büchel, T. Paladini, S. Longari, M. Carminati, S. Zanero, H. Binyamini, G. Engelberg, D. Klein, G. Guizzardi, M. Caselli, A. Continella, M. van Steen, A. Peter, and T. van Ede, "SoK: Automated TTP Extraction from CTI Reports – Are We There Yet?" in *34th USENIX Security Symposium (USENIX Security 25)*, 2025, pp. 4621–4641.
- [18] R. Fayyazi, R. Taghdimi, and S. J. Yang, "Advancing TTP Analysis: Harnessing the Power of Large Language Models with Retrieval Augmented Generation," in *2024 Annual Computer Security Applications Conference Workshops (ACSAC Workshops)*, Dec. 2024, pp. 255–261.
- [19] M. Elam, D. Mink, S. S. Bagui, R. Plenkers, and S. C. Bagui, "Introducing UWF-ZeekData24: An Enterprise MITRE ATT&CK Labeled Network Attack Traffic Dataset for Machine Learning/AI," Apr. 2025.
- [20] "Nlp-based cyber security dataset," <https://www.kaggle.com/datasets/hussainsheikh03/nlp-based-cyber-security-dataset>, 2024, kaggle dataset.
- [21] "Cve dataset," <https://www.kaggle.com/datasets/casimireffect/cve-dataset>, 2024, kaggle dataset.
- [22] Q. Cheng, L. Chen, Z. Hu, J. Tang, Q. Xu, and B. Ning, "A novel prompting method for few-shot NER via LLMs," *Natural Language Processing Journal*, vol. 8, p. 100099, Sep. 2024.
- [23] B. Agarwal, I. Joshi, and V. Rojkova, "Think Inside the JSON: Reinforcement Strategy for Strict LLM Schema Adherence," Feb. 2025.
- [24] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, and Q. Li, "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '24. New York, NY, USA: Association for Computing Machinery, Aug. 2024, pp. 6491–6501.