General Comments

This document summarizes the revisions made since the previous submission. We have carefully addressed all reviewer feedback, corrected typos and misrepresentations, and improved the overall writing and structure for better clarity. The Experimental Setup (Section 4) has been improved with additional baselines and the Results and Analysis (Section 5) have been updated accordingly with the inclusion of new analysis, 'Error Analysis' that presents the distribution of error types across six benchmark datasets, providing a clearer understanding of our findings and highlighting the bottlenecks.

Revising Title We have revised our title in order to better represent our work and the contributions of our work. The title thus has been revised from "Adaptive Prompting: A Dynamic Approach to Temporal Table Reasoning" to "No Universal Prompt: Unifying Reasoning through Adaptive Prompting for Temporal Table Reasoning".

Addition of Structural, Temporal and Agentic Baselines : Two reviewers raised concerns about the baselines used to compare our model SEAR. Previously, we evaluated SEAR only against single-step baselines such as Chain-of-Thought (CoT) (Wei et al. 2022) , Evidence Extraction (EE), Decomposed Prompting (Decomp) (Khot et al. 2023) , Faithful COT (F-COT) (Lyu et al. 2023) and Program-of-Thought (POT) (Chen et al. 2023). To address this, we have now introduced three structural baselines: Self-Discover (Zhou et al. 2024), Self-Ask (Press et al. 2023), and Plan & Solve (Wang et al. 2023a) to compare SEAR with approaches that use multi-step and selection reasoning. We also added two temporal baselines: C.L.E.A.R (Deng et al. 2025) and Narration of Thought (NoT) (Zhang, Beauchamp, and Wang 2024) to evaluate SEAR's performance on temporal reasoning tasks. Finally, we included three agentic baselines: Self-Consistency Prompting (SCP) (Wang et al. 2023b), Tree of Thought (ToT) (Yao et al. 2023), and Graph of Thought (GoT) (Besta et al. 2023) to further assess SEAR against advanced agentic and multi-path reasoning approaches. All baselines evaluated are listed in Table 4 and their respective Hybrid Correctness Score (HCS) are captured in Table 6,7,8.

Analysis of New Baselines : We have added results for the new baselines in Tables 6, 7, and 8, which report the HCS scores. Additionally, we have highlighted the performance of these newly added baselines in Lines 375-388. From these results, we observe that prompting strategies such as Self-Discover, NoT, C.L.E.A.R., and GoT perform comparatively well; however, they struggle to generalize, with performance dropping on datasets such as Multi-HierTT, Squall, and HybridQA.

Empirically evaluating information retained after table refactoring One reviewer raised a concern about potential information loss during the table refactoring step. To address this, we empirically evaluated the refactored tables using the AutoQA metric (Jain, Marzoca, and Piccinno 2024) and reported their information accuracy in Table 9 (Lines 389-400). For all tables more than 95% of information is retained except for Multi_HierTT, HybridQA and Squall with retained information above 85%.

Conducted Error Analysis : In response to the reviewers' request to analyze negative scenarios, we manually inspected 150 model errors (25 random queries from each of six datasets: HybridQA, HiTabs, Multi-HierTT, TAT-QA, FetaQA, and WikiTable Questions) and labeled each as (i) evidence extraction, (ii) reasoning, or (iii) Python-code errors. The resulting distribution (Figure 2) shows that evidence extraction dominates in five of the six datasets, accounting for 64-92% of errors. This confirms that most failures occur before any logical or numerical reasoning is attempted, with the main bottleneck being the extraction of relevant information. The one exception is WikiTable Questions, where large table aggregation and incomplete parsing make code errors the largest share (44%). For TAT-QA, common issues included temporal mismatches (e.g., Q1 vs. FY) and incorrect unit normalization (e.g., billions vs. millions). In datasets such as HybridQA, Multi-HierTT, and HiTabs, correctly identifying headers or row/column information proved challenging. These findings highlight the primary bottlenecks in the pipeline and motivate future work on stronger retrieval mechanisms.

Restructuring of Appendix : In the previous version, reviewers found the Appendix difficult to navigate. In the revised version, we have organized it more clearly, added an introductory paragraph describing its contents, and included explicit references to its figures and tables for easier navigation. We renamed Appendix A from "Example Appendix" to "Prompt Examples" and updated it to include figures representing the 3-step SEAR, SEAR_Unified, Evaluation, and Refactoring prompts (Lines 875-879). In Appendix B, we added the definitions and results for Relaxed Exact Match Score (REMS) and Contextual Answer Evaluation (CAE). Appendix B also includes tables presenting the reasoning path distribution for LLaMA and Gemini. Appendix C includes the complete table and context used to create Figure 1. Appendix D provides a detailed error analysis for six datasets, extending the analysis presented in Section 5. Finally, Appendix E contains descriptions of each dataset used in the study.

Additional Refactoring Updates We also improved several other sections to strengthen our explanations. These enhancements were made to ensure clarity, rigor, and a more comprehensive understanding of our work, thereby providing stronger support for our conclusions.

- **Refactored dataset information into a single table** In the previous version, each dataset and its question count were described in separate paragraphs. In the revised version, we consolidated this information into Table 3, while moving the detailed descriptions to Appendix E.
- **Baselines at glance** Similar to the datasets, we summarized all evaluated baselines in Table 4, grouping them by category. This improves readability and highlights the breadth of our experimental coverage.
- **Improved Introduction** We revised the Introduction to make it more concise and clearly convey the motivation

for our work. To illustrate the core idea, we added a carpenter analogy (Lines 86-97) that provides an intuitive explanation. Also, included Table 1 which showcases diversity of dataset used in the study.

- **Added 3-Step SEAR Prompts** We added Figures 3, 4, and 5 to illustrate the three steps of SEAR in order. These figures include the corresponding prompts and sample responses for each step, making it easier for others to understand and apply the prompts, thereby promoting reproducibility.

References

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Podstawski, M.; Niewiadomski, H.; Nyczyk, P.; and Hoefler, T. 2023. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. arXiv:2308.09687.

Chen, W.; Ma, X.; Wang, X.; and Cohen, W. W. 2023. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. arXiv:2211.12588.

Deng, I.; Dixit, K.; Roth, D.; and Gupta, V. 2025. Enhancing Temporal Understanding in LLMs for Semi-structured Tables. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Findings of the Association for Computational Linguistics: NAACL 2025*, 4936–4955. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-195-7.

Jain, P.; Marzoca, A.; and Piccinno, F. 2024. STRUCTSUM Generation for Faster Text Comprehension. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7876–7896. Bangkok, Thailand: Association for Computational Linguistics.

Khot, T.; Trivedi, H.; Finlayson, M.; Fu, Y.; Richardson, K.; Clark, P.; and Sabharwal, A. 2023. Decomposed Prompting: A Modular Approach for Solving Complex Tasks. arXiv:2210.02406.

Lyu, Q.; Havaldar, S.; Stein, A.; Zhang, L.; Rao, D.; Wong, E.; Apidianaki, M.; and Callison-Burch, C. 2023. Faithful Chain-of-Thought Reasoning. In Park, J. C.; Arase, Y.; Hu, B.; Lu, W.; Wijaya, D.; Purwarianti, A.; and Krisnadhi, A. A., eds., *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 305–329. Nusa Dua, Bali: Association for Computational Linguistics.

Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N.; and Lewis, M. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5687–5711. Singapore: Association for Computational Linguistics.

Wang, L.; Xu, W.; Lan, Y.; Hu, Z.; Lan, Y.; Lee, R. K.-W.; and Lim, E.-P. 2023a. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2609–2634. Toronto, Canada: Association for Computational Linguistics.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023b. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601.

Zhang, X. F.; Beauchamp, N.; and Wang, L. 2024. Narrative-of-Thought: Improving Temporal Reasoning of Large Language Models via Recounted Narratives. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida: Association for Computational Linguistics.

Zhou, P.; Pujara, J.; Ren, X.; Chen, X.; Cheng, H.-T.; Le, Q. V.; Chi, E. H.; Zhou, D.; Mishra, S.; and Zheng, H. S. 2024. Self-Discover: Large Language Models Self-Compose Reasoning Structures. arXiv:2402.03620.