ARN: Analogical Reasoning on Narratives*

Zhivar Sourati[1,2][0000−0003−2129−6165], Filip Ilievski[3][0000−0002−1735−0686], Pia Sommerauer[4][0000−0003−3593−1465], and Yifan Jiang[1,2][0000−0003−2851−9210]

[1] Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA

[2] Department of Computer Science, University of Southern California, Los Angeles, CA, USA

{souratih,yifjia}@isi.edu

[3] Department of Computer Science, Vrije Universiteit Amsterdam, The Netherlands

f.ilievski@vu.nl

[4] Computational Linguistics & Text Mining Lab, Vrije Universiteit Amsterdam, The Netherlands

pia.sommerauer@vu.nl

1 Introduction

Large language models (LLMs) have already started to outperform human baselines on various tasks, including some that require mere language understanding [16] and others that also require reasoning [18, 21]. However, one of the main issues remaining to be solved is their generalizability to new situations or domains [10, 4]. A critical cognitive skill that enables generalization in humans is analogical reasoning [12, 5, 3]. With analogical reasoning, humans can perceive, discern, and utilize the similarities between situations or events based on (systems of) relations rather than surface similarities [6, 1]. Due to the importance of this ability for AI systems, many studies have created analogical reasoning benchmarks for language and visual models [17, 20, 15, 9, 14]. An opportunity arises to ground AI benchmarking in analogical reasoning theories in cognitive psychology. We note two gaps in this direction: larger-scale benchmarks have commonly focused on word-based proportional analogies of the form A:B::C:D [11], whereas cognitively grounded benchmarks with longer texts are usually limited in size but richer in terms of theoretical depth and the complexity of the captured analogies [8, 19]. The narrow scientific context of these benchmarks hinders the scientific insight into LLMs' analogical reasoning in more common, daily situations.

2 Methodology

Our paper's methodological contributions are twofold. First, we design a **comprehensive theory-grounded framework** (Figure 1) that extracts analogies from narratives by operationalizing the link between existing analogical and

* This paper [13] was accepted for publication in the Transactions of the Association for Computational Linguistics (TACL) journal on 29/04/2024.

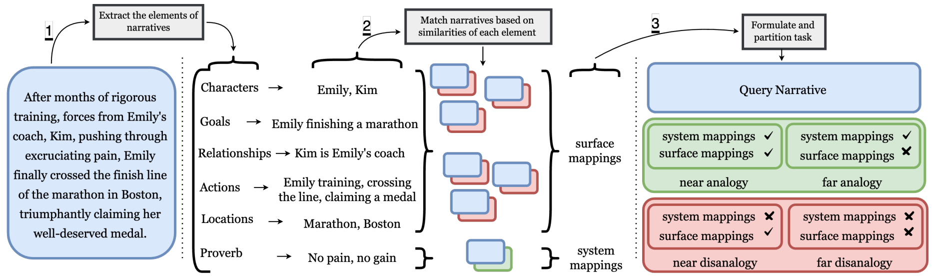

Fig. 1. The *ARN* framework for evaluating analogical reasoning on narratives.

narratology theories. Then, we introduce a binary QA task and benchmark: **A**nalogical **R**easoning over **N**arratives (*ARN*) containing 1.1k triples of query narratives, analogies, and distractors. Drawing on cognitive theories of analogy [2, 7], we include **system mappings**, **surface mappings**, and their interactions in *ARN*, which help us characterize analogical reasoning abilities of LLMs in four different scenarios with varying levels of difficulty. The four levels are designed to test the recognition of near/far analogs in the presence of near/far disanalogs. ARN employs proverbs as the basis for system mappings and contrasts them with surface-level mappings of overlapping characters, goals, actions, locations, and relations. Utilizing **proverbs** in system mappings as distilled forms of human wisdom, complex relationships, and moral lessons, and **narratives** as the primary medium in which people communicate, *ARN* facilitates extrapolating the benchmarking findings to daily analogical reasoning tasks that LLMs could be used for.

3 Results and Discussion

Human performance remains consistently high on both near and far analogies. Evaluating multiple LLMs on *ARN* in a zero-shot setting suggests that while models can recognize near analogies, their analogical reasoning performance degrades (by 35 absolute points on average) when detecting far analogies characterized by the absence of surface mappings. This trend also holds for GPT4.0, performing best on average but dropping to a below-random performance on detecting far analogies in a zero-shot setting. Few-shot prompting with Chain-of-Thought reasoning enhances models' performance in far analogies while being detrimental to solving near analogies. Overall, we show that LLMs' analogical reasoning over narratives lags behind humans, especially on far analogies, which motivates further research on devising computational analogical reasoners on narratives. Inspired by these findings, we plan to explore personalization in analogical reasoning, investigating how models tailor analogy generation to individual users and contexts. We intend to improve AI systems' relevance, adaptability, and impact in real-world applications. ARN and the entire code of this analysis are publicly released to support such endeavors at https://bit.ly/3xVTjbL.

References

1. Gentner, D., Smith, L., Ramachandran, V.: Analogical reasoning, 2012. Encyclopedia of Human Behavior, 2nd ed., VS Ramachandran, ed., Elsevier, Oxford, UK pp. 130–136 (2012)
2. Gentner, D.: Structure-mapping: A theoretical framework for analogy. Cognitive science **7**(2), 155–170 (1983)
3. Goswami, U.: Analogical reasoning in children. Psychology Press (2013)
4. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't stop pretraining: Adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964 (2020)
5. Hofstadter, D.R.: Analogy as the core of cognition. The analogical mind: Perspectives from cognitive science pp. 499–538 (2001)
6. Holyoak, K.J.: 234 Analogy and Relational Reasoning. In: The Oxford Handbook of Thinking and Reasoning. Oxford University Press (03 2012). https://doi.org/10.1093/oxfordhb/9780199734689.013.0013, https://doi.org/10.1093/oxfordhb/9780199734689.013.0013
7. Holyoak, K., Thagard, P.: Mental Leaps: Analogy in Creative Thought. A Bradford book, MIT Press (1996), https://books.google.com/books?id=8ZRHYv59154C
8. Ichien, N., Lu, H., Holyoak, K.J.: Verbal analogy problem sets: An inventory of testing materials. Behavior research methods **52**, 1803–1816 (2020)
9. Jiayang, C., Qiu, L., Chan, T.H., Fang, T., Wang, W., Chan, C., Ru, D., Guo, Q., Zhang, H., Song, Y., et al.: Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. arXiv preprint arXiv:2310.12874 (2023)
10. Li, C., Tian, Y., Zerong, Z., Song, Y., Xia, F.: Challenging large language models with new tasks: A study on their adaptability and robustness. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics ACL 2024. pp. 8140–8162. Association for Computational Linguistics, Bangkok, Thailand and virtual meeting (Aug 2024), https://aclanthology.org/2024.findings-acl.485
11. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 746–751. Association for Computational Linguistics, Atlanta, Georgia (Jun 2013), https://aclanthology.org/N13-1090
12. Penn, D.C., Holyoak, K.J., Povinelli, D.J.: Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. Behavioral and brain sciences **31**(2), 109–130 (2008)
13. Sourati, Z., Ilievski, F., Sommerauer, P., Jiang, Y.: Arn: Analogical reasoning on narratives (2024), https://arxiv.org/abs/2310.00996
14. Sultan, O., Bitton, Y., Yosef, R., Shahaf, D.: Parallelparc: A scalable pipeline for generating natural-language analogies. arXiv preprint arXiv:2403.01139 (2024)
15. Ushio, A., Espinosa Anke, L., Schockaert, S., Camacho-Collados, J.: BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 3609–3624. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.280, https://aclanthology.org/2021.acl-long.280

16. Wang, Z., Xie, Q., Ding, Z., Feng, Y., Xia, R.: Is chatgpt a good sentiment analyzer? a preliminary study. arXiv preprint arXiv:2304.04339 (2023)
17. Webb, T., Holyoak, K.J., Lu, H.: Emergent analogical reasoning in large language models. Nature Human Behaviour **7**(9), 1526–1541 (2023)
18. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
19. Wharton, C., Holyoak, K., Downing, P., Lange, T., Wickens, T., Melz, E.: Below the surface: Analogical similarity and retrieval competition in reminding. Cognitive Psychology **26**(1), 64–101 (1994). https://doi.org/https://doi.org/10.1006/cogp.1994.1003, https://www.sciencedirect.com/science/article/pii/S0010028584710036
20. Wijesiriwardene, T., Wickramarachchi, R., Gajera, B., Gowaikar, S., Gupta, C., Chadha, A., Reganti, A.N., Sheth, A., Das, A.: ANALOGICAL - a novel benchmark for long text analogy evaluation in large language models. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 3534–3549. Association for Computational Linguistics, Toronto, Canada (Jul 2023), https://aclanthology.org/2023.findings-acl.218
21. Zhou, X., Zhang, Y., Cui, L., Huang, D.: Evaluating commonsense in pre-trained language models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 9733–9740 (2020)