Zhivar Sourati

+1 (818) 877 3590 | souratih@usc.edu, zhivarsourati@gmail.com | zhpinkman.github.io | Google Scholar

Education

Ph.D. Computer Science | University of Southern California, CA, US | 2022 – Present (expected: May 2027)

Reasoning in Language Models, Analogical Reasoning, Prototype-based Reasoning, Natural Language Processing, Mechanistic Interpretability, Computational Social Sciences, and Large Language Models (LLMs) (GPA: 4.0 / 4.0)

B.Sc. Computer Engineering | University of Tehran, Tehran, Iran | 2017 - 2021

Thesis: Comparison of the Performance of Permutation and Randomization Tests on Graphs (GPA: 4.0 / 4.0)

Research & Work Experience

Research Assistant | USC / ISI | 2022 – Present

- Understand, evaluate, and improve LMs' reasoning through cognitively inspired and explainable methods like analogical, prototype-based reasoning, and mechanistic interpretability methods, e.g., utilizing case-based and analogical reasoning for logical fallacy detection, creation of an analogical reasoning benchmark on narratives, improving LMs' robustness using prototype-based reasoning, and using mechanistic interpretability to decipher analogical reasoning in LMs
- Study the effects and applications of LLMs in social psychology, particularly personality linguistic markers

Research Intern | Genentech | 2024

- Worked as a knowledge graph and large language models intern on KG-aware biomedical representations

NLP Research Assistant | Zurich University of Applied Sciences | 2021 – 2022

- Researched automatic extractive and abstractive summarization techniques using Transformers and knowledge graphs on dialogues and their transcripts
- Conducted hate speech analysis and prediction on Twitter timelines

Research Assistant | University of Tehran | 2020 – 2021

- Examined machine learning techniques in social network analysis and non-parametric (i.e., permutation) tests on complex networks (e.g., global trade network) using models like Exponential Random Graph Models (ERGMs)
- Investigated the propagation and prevalence of COVID-related topics on social media (e.g., Twitter) both from a graph and NLP perspective
- Studied different models and concepts in reinforcement learning, such as n-armed bandits, On-/Off-policy methods, and investigated their practical use cases, e.g., analyzing the monetary value of time using Prospect Theory

NLP Research Intern | TeIAS | 2020

- Studied SOTA and common models, datasets, and tasks in NLP, such as NER and QA chatbots
- Compiled the Datasets for Farsi (Persian) Natural Language Processing

Publications & Presentations

- Sourati, Z., Ozcan, M., Karimi, F., McDaniel, C., Ziabari, A., Trager, J., Wen, N., Tak, A., Morstatter, F., & Dehghani, M. (2024). Secret Keepers: The Impact of LLMs on Linguistic Markers of Personal Traits. arXiv preprint arXiv:2404.00267. (Under review)

- Ahrabian, K., Sourati, Z., Sun, K., Zhang, J., Jiang, Y., Morstatter, F., & Pujara, J. (2024). The Curious Case of Nonverbal Abstract Reasoning with Multi-Modal Large Language Models. arXiv preprint arXiv:2401.12117. (Accepted to COLM 2024)
- Jiang, Y., Zhang, J., Sun, K., Sourati, Z., Ahrabian, K., ... & Pujara, J. (2024). MARVEL: Multidimensional Abstraction and Reasoning through Visual Evaluation and Learning. *arXiv preprint arXiv:2404.13591*. (Accepted to NeurIPS 2024)
- Sourati, Z., Ilievski, F., Sommerauer, P., & Jiang, Y. (2024). ARN: Analogical Reasoning on Narratives. *Transactions of the Association for Computational Linguistics*, *12*, 1063-1086. (Presented at EMNLP 2024)
- Deshpande, D., Sourati, Z., Ilievski, F., & Morstatter, F. (2024). Contextualizing Argument Quality Assessment with Relevant Knowledge. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 316–326, Mexico City, Mexico. Association for Computational Linguistics.
- Sourati, Z., Deshpande, D., Ilievski, F., Gashteovski, K., & Saralajew, S. (2024). Robust Text Classification: Analyzing Prototype-Based Networks. arXiv preprint arXiv:2311.06647. (Accepted to EMNLP 2024)
- Jiang, Y., Ilievski, F., Ma, K., & Sourati, Z. (2023). BRAINTEASER: Lateral Thinking Puzzles for Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 14317–14332, Singapore. Association for Computational Linguistics.
- Thakur, A. K., Ilievski, F., Sandlin, H. Â., Sourati, Z., Luceri, L., Tommasini, R., & Mermoud, A. (2023). Explainable Classification of Internet Memes. In NeSy (pp. 395-409).
- Sourati, Z., Ilievski, F., Sandlin, H.-Â., & Mermoud, A. (2023). Case-based reasoning with language models for classification of logical fallacies. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI ’23.
- Sourati, Z., Venkatesh, V. P. P., Deshpande, D., Rawlani, H., Ilievski, F., Sandlin, H. Â., & Mermoud, A. (2023). Robust and explainable identification of logical fallacies in natural language arguments. Knowledge-Based Systems, 266, 110418.
- Thakur, A., Ilievski, F., Sandlin, H., Mermoud, A., Sourati, Z., Luceri, L., & Tommasini, R. (2023). Multimodal and Explainable Internet Meme Classification. In AI for Social Good (AI4SG) at AAAI-23. Retrieved from https://amulyayadav.github.io/AI4SG2023/.
- ShabaniMirzaei, T., Chamani, H., Abaskohi, A., Sourati, Z., & Bahrak, B. (2023). A large-scale analysis of Persian Tweets regarding Covid-19 vaccination. Social Network Analysis and Mining, 13(1), 148.
- Sourati, Z., Sabri, N., Chamani, H., & Bahrak, B. (2022). Quantitative analysis of fanfictions’ popularity. Social Network Analysis and Mining, 12(1), 42.
- Setayesh, A., Sourati, Z., & Bahrak, B. (2022). Analysis of the global trade network using exponential random graph models. Applied Network Science, 7(1), 38.
- Sourati, Z., von Däniken, P., Cieliebak, M. (2022). Ukraine-Russia - First insights into recent Twitter posts about this conflict. SwissText Conference; June 2022; Lugano, Switzerland. (Poster presentation)
- Von Däniken, P., Sourati, Z., Tuggener, D. (2022). Hateful Social Media Users - Can we predict their behavior? SwissText Conference; June 2022; Lugano, Switzerland. (Poster presentation)
- Chamani, H., Sourati, Z., & Bahrak, B. (2021, October). An Overview of Regression Methods in Early Prediction of Movie Ratings. In *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)* (pp. 1-6). IEEE.

Skills

- Python, R, C, C++, SQL, Java, Typescript, SPARQL
- PyTorch, Hugging Face, Weights & Biases, Keras, Stata, Linux, Spring, Angular, Git, MongoDB, Node js., MySQL, Neo4j