Reasoning with Natural Language Explanations

Marco Valentino[1], **André Freitas**[1,2,3]
[1]Idiap Research Institute, Switzerland,
[2]Department of Computer Science, University of Manchester, UK
[3]National Biomarker Centre, CRUK-MI, University of Manchester, UK
first.last@idiap.ch

Abstract

Explanation constitutes an archetypal feature of human rationality, underpinning learning, and generalisation, and representing one of the media supporting scientific discovery and communication. Due to the importance of explanations in human reasoning, an increasing amount of research in Natural Language Inference (NLI) has started reconsidering the role that explanations play in learning and inference, attempting to build explanation-based NLI models that can effectively encode and use natural language explanations on downstream tasks. Research in explanation-based NLI, however, presents specific challenges and opportunities, as explanatory reasoning reflect aspects of both material and formal inference, making it a particularly rich setting to model and deliver complex reasoning. In this tutorial, we provide a comprehensive introduction to the field of explanation-based NLI, grounding this discussion on the epistemological-linguistic foundations of explanations, systematically describing the main architectural trends and evaluation methodologies which can be used to build systems which are capable of explanatory reasoning[1].

1 Introduction

Building systems that can understand and explain the world is a long-standing goal for *Artificial Intelligence (AI)* (Miller, 2019; Mitchell et al., 1986; Thagard and Litt, 2008). The ability to explain, in fact, constitutes an archetypal feature of human rationality, underpinning communication, learning, and generalisation, as well as one of the mediums enabling scientific discovery and progress through the formulation of explanatory theories (Lombrozo, 2012; Salmon, 2006; Kitcher, 1989; Deutsch, 2011).

Due to the importance of explanation in human reasoning, an increasing amount of work has started reconsidering the role that explanation plays in learning and inference with natural language (Camburu et al., 2018; Yang et al., 2018; Rajani et al., 2019; Jansen et al., 2018). In contrast to the existing end-to-end paradigm based on Deep Learning, explanation-based NLI focuses on developing and evaluating models that can address downstream tasks through the explicit construction of a *natural language explanation* (Dalvi et al., 2021; Jansen et al., 2016; Wiegreffe and Marasović, 2021; Stacey et al., 2022). In this context, explanation is seen as a potential solution to mitigate some of the well-known limitations in neural-based NLI architectures (Thayaparan et al., 2020), including the susceptibility to learning via shortcuts, the inability to generalise out-of-distribution, and the lack of interpretability (Guidotti et al., 2018; Biran and Cotton, 2017; Geirhos et al., 2020; Lewis et al., 2021; Sinha et al., 2021; Schlegel et al., 2020).

Research in explanation-based NLI, however, presents several fundamental challenges (Valentino and Freitas, 2024). First, the applied methodologies are still poorly informed by theories and accounts of explanations (Salmon, 2006; Woodward and Ross, 2021). This gap between theory and practice poses the risk of slowing down progress, missing the opportunity to formulate clearer hypotheses on the inferential properties of natural language explanations and define systematic evaluation methodologies (Camburu et al., 2020; Jansen et al., 2021; Atanasova, 2024). Second, explanation-based NLI models still lack robustness, control, and scalability for real-world applications. In particular, existing approaches suffer from several limitations when composing explanatory reasoning chains and performing abstraction for NLI in complex domains (Khashabi et al., 2019; Valentino et al., 2022a).

In this tutorial, we will provide a comprehensive introduction to explanatory reasoning in the context of NLI, by systematically categorising and surveying explanation-supporting benchmarks, ar-

[1]Tutorial website: https://sites.google.com/view/reasoning-with-explanations

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 25–31
November 12–16, 2024.

chitectures, and research trends. Specifically, we will present how the understanding of explanatory inference have evolved in recent years, together with the emerging methodological and modelling strategies. In parallel, we will attempt to provide an epistemological-linguistic characterisation of natural language explanations reviewing the main theoretical accounts (Valentino and Freitas, 2024; Salmon, 2006) to derive a fresh perspective for future work in the field.

2 Description

This section outlines the content of the tutorial.

2.1 Epistemological-Linguistic Foundations

One of the main objectives of the tutorial is to provide a theoretically grounded foundation for explanation-based NLI, investigating the notion of explanation as a language and inference scientific object of interest, from both an *epistemological* and *linguistic* perspectives (Valentino and Freitas, 2024; Salmon, 2006; Jansen et al., 2016).

To this end, we will present a systematic survey of the contemporary discussion in Philosophy of Science around the notion of a scientific explanation, attempting to shed light on the nature and function of explanatory arguments and their constituting elements. Here, we will critically review the main accounts of explanations, including the deductive-nomological and inductive-statistical account (Hempel and Oppenheim, 1948), the notion of statistical relevance and the causal-mechanical model (Salmon, 1984), and the unificationist account (Kitcher, 1989), aiming to elicit what it means to perform explanatory reasoning. Following the survey, we will focus on grounding the theoretical accounts for explanation-based NLI, attempting to identify the main feature of explanatory arguments in existing corpora of natural language explanations (Jansen et al., 2016; Xie et al., 2020; Jansen et al., 2018).

2.2 Resources & Evaluation Methods for Explanation-Based NLI

In order to build NLI models that can reason through the generation of natural language explanations it is necessary to develop systematic evaluation methodologies. To this end, The tutorial will review the main resources, benchmarks and metrics in the field (Wiegreffe and Marasovic).

Depending on the nature of the NLI problem, an explanation can include pieces of evidence at different levels of abstraction (Thayaparan et al., 2020). Traditionally, the field has been divided into *extractive* and *abstractive* tasks. In extractive NLI, the reasoning required for the explanations is derivable from the original problem formulation, where the correct decomposition of the problem contains all the necessary inference steps for the answer (Yang et al., 2018). On the other hand, abstractive NLI tasks require going beyond the surface form of the problem, where an explanation needs to account for and cohere definitions, abstract relations, which are not immediately available from the original context (Jansen et al., 2021; Thayaparan et al., 2021b).

In addition, the tutorial will review the main evaluation metrics adopted to assess the quality of natural language explanations. Evaluating the quality of explanations, in fact, is a challenging problem as it requires accounting for multiple concurrent properties. Different metrics have been proposed in the field, ranging from reference-based metrics designed to assess the alignment between automatically generated explanations and human-annotated explanations (Camburu et al., 2018; Jansen et al., 2021), and reference-free metrics designed to evaluate additional dimensions such as faithfulness (Parcalabescu and Frank, 2024; Atanasova et al., 2023), robustness (Camburu et al., 2020), logical validity (Quan et al., 2024b; Valentino et al., 2021a), and plausibility (Dalal et al., 2024).

2.3 Explanation-Based Learning & Inference

We review the key architectural patterns and modelling strategies for reasoning and learning over natural language explanations. In particular, we focus on the following paradigms:

Multi-Hop Reasoning & Retrieval-Based Models. The construction of explanations typically requires multi-hop reasoning – i.e., the ability to compose multiple pieces of evidence to support the final answer (Dalvi et al., 2021; Xie et al., 2020). Multi-hop reasoning has been largely studied in a retrieval settings, where, given an external knowledge base, the model is required to select, collect and link the relevant knowledge required to arrive at a final answer (Valentino et al., 2022a, 2021b, 2022b). Here, we will review the main retrieval-based architectures for multi-hop reasoning and explanation, highlighting some of the inherent limitations of such paradigm, including the tension between semantic drift and efficiency (Khashabi

et al., 2019).

Natural Language Explanation Generation. In parallel with retrieval approaches, NLI using generative models have been used for supporting explanatory inference (Camburu et al., 2018; Rajani et al., 2019). In this setting, early approaches leverage human-annotated natural language explanations for training generative models (Dalvi et al., 2021). Subsequently, the advent of Large Language Models (LLMs) has made it possible to elicit explanatory reasoning via specific prompting techniques and in-context learning (Wei et al., 2022; Yao et al., 2024; Zheng et al., 2023; He et al., 2024). Here, we review the main trends in the LLM-based generative paradigms, highlighting persisting limitations such as hallucinations and faithfulness (Turpin et al., 2024).

2.4 Semantic Control for Explanatory Reasoning

Controlling the explanation generation process in neural-based models is particularly critical while modelling complex reasoning tasks. In this tutorial, we will review emerging trends which combine neural and symbolic approaches to improve semantic control in the explanatory reasoning process, which can provide formal guarantees on the quality of the explanations. These methods aim to integrate the content flexibility of language models (instrumental for supporting material inferences) and a formal inference properties.

In particular, we focus on the following key methods:

Leveraging Explanatory Inference Patterns for Explanation-Based NLI. Inference patterns in explanation corpora can be leveraged to improve the efficiency and robustness of neural representations (Valentino and Freitas, 2024; Zhang et al., 2023). In particular, we will review approaches that attempt to leverage the notion of unification power in corpora of natural language explanations to improve multi-hop reasoning in a retrieval setting and alleviate semantic drift (Valentino et al., 2022a, 2021b, 2022b).

Constraint-Based Optimisation for Explanation-Based NLI. We will focus on describing neuro-symbolic methods which target encoding explicit assumptions about the structure of natural language explanations (Thayaparan et al., 2021a). Here, we will review methods performing multi-hop inference via constrained optimisation, integrating neural representations with explicit constraints via end-to-end differentiable optimisation approaches (Thayaparan et al., 2022, 2024).

Formal-Geometric Inference Controls over Latent Spaces. Covers emerging methodologies which focus on learning latent spaces with better representational properties for explanatory NLI, using language Variational Autoencoders (VAEs) for delivering better disentanglement and separability of language and inference properties (Zhang et al., 2024a,c,b,a) which support better inference control. These methods deliver an additional geometrical structure to latent spaces, aiming to deliver the vision of 'inference as latent geometry'.

LLM-Symbolic Architectures Finally, we will focus on hybrid neuro-symbolic architectures that attempt to leverage the material/content-based inference properties of LLMs for explanation generation with external symbolic approaches, which accounts for formal/logical validity refinement properties. In particular, we will review approaches that perform explanation refinement via the integration of LLMs and Theorem Provers to verify logical validity (Quan et al., 2024b,a) and additional external tools to evaluate explanation properties such as uncertainty, plausibility and coherence (Dalal et al., 2024).

3 Schedule

The tutorial will be organised according to the following timeline:

1. Introduction & Motivation (20 min.)
2. Epistemological-Linguistic Foundations (20 min.)
3. Resources & Evaluation for Explanation-Based NLI (40 min.)
4. Explanation-Based Learning & Inference (40 min.)
5. Semantic Control for Explanatory Reasoning (40 min.)
6. Synthesis, Discussion, and Q&A (20 min)

4 Breadth & Diversity

The tutorial will cover a wide spectrum of topics in different fields, ranging from Philosophy,

Machine Learning, Natural Language Processing, Knowledge Representation and Automated Reasoning. This diversity of topics will help create a rich environment in which academics from different backgrounds and cultural contexts can integrate different perspectives. The tutorial plan includes integrated open Q&A sessions and practical demonstrations.

5 Prerequisites

We do not expect attendees to be familiar with previous research on NLI and Explanatory inference. On the opposite, we intent this tutorial to be an efficient and deep onboarding into the state-of-the-art in those areas. Participants should have a general background knowledge in deep learning, including recent trends and architectures such as Large Language Models. Participants are expected to be familiar with some of the broader NLI tasks, such as Textual Entailment and Question Answering.

6 Reading List

Epistemological-Linguistic Foundations

Valentino and Freitas (2024) On the Nature of Explanation: An Epistemological-Linguistic Perspective for Explanation-Based Natural Language Inference.

Salmon (2006) Four Decades of Scientific Explanation.

Jansen et al. (2016) What's in an Explanation? Characterizing Knowledge and Inference Requirements for Elementary Science Exam.

Resources, Models and Evaluation

Wiegreffe and Marasović (2021) Teach me to Explain: A Review of Datasets for Explainable Natural Language Processing.

Thayaparan et al. (2020) A Survey on Explainability in Machine Reading Comprehension.

Zhao et al. (2024) Explainability for Large Language Models: A Survey.

Related Tutorials

Zhu et al. (2024) Explanation in the Era of Large Language Models.

Camburu and Akata (2021) Natural-XAI: Explainable AI with Natural Language Explanation.

Zhao et al. (2023) Complex Reasoning in Natural Language.

Boyd-Graber et al. (2022) Human-Centered Evaluation of Explanations.

7 Instructor information

Marco Valentino, Idiap Research Institute.[2] Marco is a postdoctoral researcher at the Idiap Research Institute, Switzerland. His research is carried out at the intersection of Natural Language Inference and Neuro-Symbolic models focusing on building systems that can reason through natural language explanations in complex domains (e.g., mathematics, science, biomedical and clinical applications, ethical reasoning). He has published papers in major AI and NLP conferences including AAAI, ACL, EMNLP, NAACL and EACL. Marco was involved in the organisation of workshops including MathNLP (EMNLP 2022 and LREC-COLING 2024), and TextGraphs (COLING 2022 and ACL 2024).

André Freitas, University of Manchester & Idiap Research Institute.[3] André Freitas leads the Neuro-symbolic AI Lab at the University of Manchester and IDIAP Research Institute. His main research interests are on enabling the development of AI methods to support abstract, flexible and controlled reasoning in order to support AI-augmented scientific discovery. In particular, he investigates how the combination of neural and symbolic data representation paradigms can deliver better models of inference. He is an active contributor to the main conferences and journals in the AI/Natural Language Processing (NLP) interface (AAAI, NeurIPs, ACL, EMNLP, EACL, COLING, TACL, Computational Linguistics), with over 100 peer-reviewed publications. He contributed to the organisation of MathNLP at EMNLP 2022 and LREC-COLING 2024. André participated in 7 tutorials, and co-organised 1 conference and 6 workshops.

Acknowledgements

This work was partially funded by the Swiss National Science Foundation (SNSF) project NeuMath (200021_204617).

[2] mailto:marco.valentino@idiap.ch
[3] mailto:andre.freitas@manchester.ac.uk

References

P Atanasova, OM Camburu, C Lioma, T Lukasiewicz, JG Simonsen, and I Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 283–294. Association for Computational Linguistics (ACL).

Pepa Atanasova. 2024. A diagnostic study of explainability techniques for text classification. In *Accountable and Explainable Methods for Complex Reasoning over Text*, pages 155–187. Springer.

Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8.

Jordan Boyd-Graber, Samuel Carton, Shi Feng, Q. Vera Liao, Tania Lombrozo, Alison Smith-Renner, and Chenhao Tan. 2022. Human-centered evaluation of explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 26–32, Seattle, United States. Association for Computational Linguistics.

Oana-Maria Camburu and Zeynep Akata. 2021. Natural-xai: Explainable ai with natural language explanations. In *International Conference on Machine Learning*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, pages 9539–9549.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.

Dhairya Dalal, Marco Valentino, Andre Freitas, and Paul Buitelaar. 2024. Inference to the best explanation in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 217–235, Bangkok, Thailand. Association for Computational Linguistics.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370.

David Deutsch. 2011. *The beginning of infinity: Explanations that transform the world*. Penguin UK.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. 2024. Using natural language explanations to improve robustness of in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13477–13499, Bangkok, Thailand. Association for Computational Linguistics.

Carl G Hempel and Paul Oppenheim. 1948. Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 2016. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2956–2965.

Peter Jansen, Kelly J Smith, Dan Moreno, and Huitzilin Ortiz. 2021. On the challenges of evaluating compositional explanations in multi-hop inference: Relevance, completeness, and expert ratings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7529–7542.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Daniel Khashabi, Erfan Sadeqi Azer, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2019. On the capabilities and limitations of reasoning for natural language understanding. *arXiv preprint arXiv:1901.02522*.

Philip Kitcher. 1989. Explanatory unification and the causal structure of the world.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008.

Tania Lombrozo. 2012. Explanation and abductive inference. *Oxford handbook of thinking and reasoning*, pages 260–276.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Tom M Mitchell, Richard M Keller, and Smadar T Kedar-Cabelli. 1986. Explanation-based generalization: A unifying view. *Machine learning*, 1(1):47–80.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089.

Xin Quan, Marco Valentino, Louise Dennis, and Andre Freitas. 2024a. Enhancing ethical explanations of large language models through iterative symbolic refinement. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–22, St. Julian's, Malta. Association for Computational Linguistics.

Xin Quan, Marco Valentino, Louise A Dennis, and André Freitas. 2024b. Verification and refinement of natural language explanations through llm-symbolic theorem proving. *arXiv preprint arXiv:2405.01379*.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Wesley C Salmon. 1984. *Scientific explanation and the causal structure of the world*. Princeton University Press.

Wesley C Salmon. 2006. *Four decades of scientific explanation*. University of Pittsburgh press.

Viktor Schlegel, Marco Valentino, André Freitas, Goran Nenadic, and Riza Theresa Batista-Navarro. 2020. A framework for evaluation of machine reading comprehension gold standards. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5359–5369.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. Unnatural language inference. 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)*, pages 7329–7346.

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Supervising model attention with human explanations for robust natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11349–11357.

Paul Thagard and Abninder Litt. 2008. Models of scientific explanation. *The Cambridge Handbook of Computational Psychology*, pages 549–564.

Mokanarangan Thayaparan, Marco Valentino, Deborah Ferreira, Julia Rozanova, and André Freitas. 2022. Diff-explainer: Differentiable convex optimization for explainable multi-hop inference. *Transactions of the Association for Computational Linguistics*, 10:1103–1119.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020. A survey on explainability in machine reading comprehension. *arXiv preprint arXiv:2010.00389*.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2021a. Explainable inference over grounding-abstract chains for science questions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1–12.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2024. A differentiable integer linear programming solver for explanation-based natural language inference. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 449–458.

Mokanarangan Thayaparan, Marco Valentino, Peter Jansen, and Dmitry Ustalov. 2021b. TextGraphs 2021 shared task on multi-hop inference for explanation regeneration. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 156–165, Mexico City, Mexico. Association for Computational Linguistics.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Marco Valentino and André Freitas. 2024. On the nature of explanation: An epistemological-linguistic perspective for explanation-based natural language inference. *Philosophy & Technology*, 37(3):88.

Marco Valentino, Ian Pratt-Hartmann, and André Freitas. 2021a. Do natural language explanations represent valid logical arguments? verifying entailment in explainable nli gold standards. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 76–86.

Marco Valentino, Mokanarangan Thayaparan, Deborah Ferreira, and André Freitas. 2022a. Hybrid autoregressive inference for scalable multi-hop explanation regeneration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11403–11411.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2021b. Unification-based reconstruction of multi-hop explanations for science questions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 200–211, Online. Association for Computational Linguistics.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2022b. Case-based abductive natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1556–1568.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Sarah Wiegreffe and Ana Marasovic. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp. *arXiv preprint arXiv:2102.12060*.

James Woodward and Lauren Ross. 2021. Scientific Explanation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Summer 2021 edition. Metaphysics Research Lab, Stanford University.

Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. 2020. Worldtree v2: A corpus of science-domain structured explanations and inference patterns supporting multi-hop inference. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5456–5473.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Yingji Zhang, Danilo Carvalho, and Andre Freitas. 2024a. Learning disentangled semantic spaces of explanations via invertible neural networks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2113–2134, Bangkok, Thailand. Association for Computational Linguistics.

Yingji Zhang, Danilo Carvalho, Marco Valentino, Ian Pratt-Hartmann, and Andre Freitas. 2024b. Improving semantic control in discrete latent spaces with transformer quantized variational autoencoders. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1434–1450, St. Julian's, Malta. Association for Computational Linguistics.

Yingji Zhang, Danilo S Carvalho, Ian Pratt-Hartmann, and Andre Freitas. 2023. Towards controllable natural language inference through lexical inference types. *arXiv preprint arXiv:2308.03581*.

Yingji Zhang, Marco Valentino, Danilo Carvalho, Ian Pratt-Hartmann, and Andre Freitas. 2024c. Graph-induced syntactic-semantic spaces in transformer-based variational AutoEncoders. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 474–489, Mexico City, Mexico. Association for Computational Linguistics.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Wenting Zhao, Mor Geva, Bill Yuchen Lin, Michihiro Yasunaga, Aman Madaan, and Tao Yu. 2023. Complex reasoning in natural language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 11–20, Toronto, Canada. Association for Computational Linguistics.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. 2023. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*.

Zining Zhu, Hanjie Chen, Xi Ye, Qing Lyu, Chenhao Tan, Ana Marasovic, and Sarah Wiegreffe. 2024. Explanation in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 19–25, Mexico City, Mexico. Association for Computational Linguistics.