SYNOPSIS
Domain Obedient Deep Learning

Soumadeep Saha

A synopsis submitted in partial fulfillment of the requirements for the degree of *Doctor of Philosophy* in Computer Science.

Supervisor: Prof. Utpal Garain

Computer Vision and Pattern Recognition Unit
Indian Statistical Institute, Kolkata

February, 2025

Abstract

Deep learning (DL), a family of data-driven artificial intelligence techniques, has shown immense promise in a plethora of applications, and it has even outpaced experts in several domains. However, unlike symbolic approaches to learning, these methods fall short when it comes to abiding by and learning from pre-existing established principles. This is a significant deficit for deployment in critical applications such as robotics, medicine, industrial automation, etc. For a decision system to be considered for adoption in such fields, it must demonstrate the ability to adhere to specified constraints, an ability missing in DL-based approaches. Exploring this problem serves as the core tenet of the dissertation.

The dissertation starts with an exploration of the abilities of conventional DL-based systems *vis-à-vis* domain coherence. A large-scale rule-annotated dataset is introduced to mitigate the pronounced lack of suitable constraint adherence evaluation benchmarks, and with its aid, the rule adherence abilities of vision systems are analyzed. Additionally, this study probes language models to elicit their performance characteristics with regard to domain consistency. Examination of these language models with interventions illustrates their ineptitude at obeying domain principles, and a mitigation strategy is proposed. This is followed by an exploration of techniques for imbuing deep learning systems with domain constraint information. Also, a comprehensive study of standard evaluation metrics and their blind spots pertaining to domain-aware performance estimation is undertaken. Finally, a novel technique to enforce constraint compliance in models without training is introduced, which pairs a search strategy with large language models to achieve cutting-edge performance.

Background

DL-based predictive systems outperform other techniques in several areas and also offer unique advantages like being able to handle myriad data types and not being reliant on expert-crafted features. However, a key hindrance to wider adoption is their *unreliability* in certain contexts. In many critical applications, automated decision-making systems *must* exhibit an ability to operate within preset boundaries or *constraints* to be considered for deployment.

It has been demonstrated that conventional DL techniques learning from data alone *do not learn* to operate within the bounds of these requisite constraints [121, 150]. Muralidhar et al. [121] showed that neural networks (NNs) trained with regular approaches disobey monotonicity and boundary constraints, and Saha et al. [150] presented similar findings for logical constraints. The problem is more severe than occasional constraint violation, and Zhang et al. [214] have noted that DL systems can fit "a random labeling of training data"–thus indicating that DL-based approaches are somewhat oblivious to underlying domain principles. This lack of domain awareness is a major deterrent to wider acceptance of DL methods in key domains.

A naive approach to dealing with this issue is to augment DL models with a rule-checking system to suppress offending predictions. Such an approach mitigates this issue somewhat, potentially at the cost of performance, but completely underutilizes the opportunity presented by the constraint information. A constraint-aware system could potentially learn from constraints to make improved predictions, infer missing details, or display enhanced robustness [14, 115]. Further, augmenting such a constraint-aware system (*soft constraint adherence*) with a rule-checker can result in *hard constraint adherence* with minimal performance disruption.

There are two pertinent areas of exploration in this regard: *learning from* and *abiding by* specified domain constraints, which is the focus of this dissertation. A **domain-obedient deep learning** system aims to leverage pre-existing domain constraints in addition to training data to improve prediction performance and better align models to domain expectations.

A related interesting area of study is rule learning or reasoning with DL-based systems [42, 137, 160], where the expected output is novel rules discovered from potentially noisy data. The constraint adherence problem, although more straightforward, has significant practical ramifications for wider applicability and is the focus of the dissertation. Awareness of domain constraints could also potentially mitigate effects of data sparsity [26, 121].

Constraints can take many forms, like numerical or logical relationships, graphs, probability distributions, or other problem-specific prior knowledge. The primary focus of this dissertation is on logical constraints. Since graph-based constraints are also typically decomposable as a set of logical constraints, a general framework for incorporating *first-order logic* (FOL) rules into DL systems would address the challenges posed by the former. Dash et al. [26] point out that "Logic is not differentiable", and addressing logical coherence poses a challenge when working in the standard DL framework, which is reliant on gradient-based optimization.

Previous studies in this area have explored techniques like modifying losses, architectures, or transforming datasets [14, 26]. The NN *architecture* employed to address a problem has a strong influence on constraint adherence. For example, consider convolutional neural networks (CNNs), which respect translation equivariance and locality constraints (spatially close pixels are semantically related), or *graph neural networks* (GNNs), which explicitly model node relationships. There have been more explicit capitalizations of this general idea, like the KBANN approach [181], which derives the structure of the NN from domain knowledge expressed as a set of propositional rules. The work by Xie et al. [205] advances a system to incorporate symbolic knowledge expressed as graphs in a GNN to improve generated embeddings. Li and Srikumar [101] proffer adding connections to the NN based on domain knowledge expressed as FOL rules.

The classical approach to *modifying losses* is to introduce auxiliary objectives penalizing incoherent predictions [112, 121, 166, 206]. Diligenti et al. [32] put forth a system to translate FOL rules to fuzzy constraints, which are then employed as penalty terms. Melacci et al. [115] rephrased domain constraints as polynomials employing continuous logics and transformed the adherence problem into an optimization problem with the polynomials serving as auxiliary losses. Melacci et al. [115] and Sheatsley et al. [165] demonstrated improved adversarial robustness with their techniques. Hu et al. [74] suggested an iterative distillation [71] technique to incorporate logical constraints.

Dataset transformations involve including background knowledge-based relational or logical features extracted from the data alongside the data [48, 97, 207]; however, when considering constraint adherence, the customary approach is to augment the dataset with examples following criteria established by domain rules. As an example, Bjerrum [13] proposes a methodology where they augment the training dataset with synthetic samples that are filtered based on domain constraints to reinforce learning of these constraints. To imbue constraints on input features, data augmentation with constraint-invariant perturbations has also been explored [118, 185].

Despite several promising forays towards logical constraint adherence, a general framework for incorporating logical constraints into DL systems remains elusive.

Dissertation Outline and Contributions

The dissertation starts (Chapter 1) with a summary of requisite background pertaining to DL techniques, architectures, language modeling, and related prior work in this area. This is followed by *four* contributory chapters.

Chapter 2 *(Do vision systems learn rules?)* focuses on DL systems' ability to adhere to domain rules for vision tasks. It introduces a new dataset to combat the lack of large-scale rule-annotated datasets and, with its assistance, demonstrates the constraint adherence deficit displayed by state-of-the-art (SoTA) DL-based computer vision (CV) techniques. Further analysis demonstrates key limitations of vision models in performing crucial tasks like counting and localization.

Chapter 3 *(Faithful Language Modeling)* analyzes language models (LMs) and points out critical deficiencies they exhibit in adhering to domain expectations with the aid of interventions during training and inference. A notion of semantic faithfulness is introduced, which demands that LMs' answers should change in response to withholding of relevant context, and it is empirically demonstrated that LMs do not abide by this requirement. Further, an intervention-based training strategy is proposed that alleviates this effect.

Chapter 4 *(Domain-aware Learning and Evaluation)* puts forth a technique called Domain Obedient Self-supervised Training (DOST) to incorporate logical constraint information into DL systems. DOST leverages domain rules alongside data to disincentivize incoherent predictions and improve predictive performance. This algorithm is more data-efficient and results in models that show enhanced *trustworthiness*. Additionally, chapter 4 tackles model evaluation and points out issues with a domain-blind approach to evaluation. A framework for constructing a metric that takes domain knowledge into account is proposed and exemplified with a real-life medical use case.

Chapter 5 *(Constrained Inference)* explores inference with large language models (LLMs) in a constrained setting. In particular, crosswords–a word puzzle featuring rich linguistic requirements alongside constraint satisfaction–are analyzed. Various subtasks of this problem are examined with an emphasis on isolating key areas of weaknesses for LLMs. Further, a new algorithm is presented, which uses in-context learning paired with a search strategy to successfully solve crosswords. This chapter presents several results that improve upon previously reported SoTA results by a factor of 2-3.

Chapter 6 consists of some concluding remarks. We highlight that the inability of DL-based systems to leverage established domain-specific principles is inexpedient. An approach to learning that acts within the framework of laid-out rules and learns from them is a key objective of DL research. The findings in the dissertation are a step towards a general framework for domain-obedient learning systems, paving the way for encouraging future research opportunities. In particular, agentic systems reasoning toward achieving goals could greatly benefit from domain-knowledge guidance, and future work tackling this challenge would be compelling.

Domain knowledge-aware DL also promises to be a major milestone towards scientific applications of deep learning. In this vein, we undertook a preliminary study attempting to address a problem arising from cosmology, where the known laws of physics serve as constraining information. Results from this study are presented in **Appendix A**.

Publications

Following is a list of publications pertinent to the dissertation.

- *MedTric : A clinically applicable metric for evaluation of multi-label computational diagnostic systems*. *Journal*
  Soumadeep Saha, Utpal Garain, Arijit Ukil, Arpan Pal, Sundeep Khandelwal.
  PLOS ONE, 18(8), 1–19; 2023. [`10.1371/journal.pone.0283895`] [148].

- *Analyzing Semantic Faithfulness of Language Models via Input Intervention on Question Answering*. *Journal*
  Akshay Chaturvedi, Swarnadeep Bhar, **Soumadeep Saha**, Utpal Garain, Nicholas Asher.
  Computational Linguistics, 50(1), 119-155; 2024. [`10.1162/coli_a_00493`] [18].

- *VALUED - Vision and Logical Understanding Evaluation Dataset*. *Journal*
  Soumadeep Saha, Saptarshi Saha, Utpal Garain.
  Journal of Data-centric Machine Learning Research (DMLR), (13):1–18; 2024. MLR press [150].

- *LADDER: Revisiting the Cosmic Distance Ladder with Deep Learning Approaches and Exploring Its Applications*. *Journal*
  Rahul Shah, **Soumadeep Saha**, Purba Mukherjee, Utpal Garain, Supratik Pal.
  The Astrophysical Journal Supplement Series (ApJS), 273(2), 1–27; 2024. The American Astronomical Society. [`10.3847/1538-4365/ad5558`] [162].

- *Language Models are Crossword Solvers*. *Conference*
  Soumadeep Saha, Sutanoya Chakraborty, Saptarshi Saha, Utpal Garain.
  Accepted to the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics [NAACL 2025 (main)].
  Pre-print: [`https://arxiv.org/abs/2406.09043`] [147].

- *DOST–Domain Obedient Self-supervision for Trustworthy Multi Label Classification with Noisy Labels*. *Workshop*
  Soumadeep Saha, Utpal Garain, Arijit Ukil, Arpan Pal, Sundeep Khandelwal.
  Proceedings of the 8th International Workshop in Health Intelligence, The Association for

the Advancement of Artificial Intelligence (AAAI), 2024.
AI for Health Equity and Fairness: Leveraging AI to Address Social Determinants of Health, 117–127; 2024. Springer Nature Switzerland. [10.1007/978-3-031-63592-2_10] [149].

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

[2] Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

[3] Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D Clifford, and Matthew A Reyna. 2021. Classification of 12-lead ECGs: the PhysioNet/computing in cardiology challenge 2020. *Physiological Measurement*, 41(12):124003.

[4] Rami Aly, Steffen Remus, and Chris Biemann. 2019. Hierarchical multi-label classification of text with capsule networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 323–330, Florence, Italy. Association for Computational Linguistics.

[5] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 39–48. IEEE Computer Society.

[6] Giovanni Angelini, Marco Ernandes, and Marco Gori. 2005. Webcrow: A web-based crosswords solver. In *Intelligent Technologies for Interactive Entertainment*, pages 295–298, Berlin, Heidelberg. Springer Berlin Heidelberg.

[7] Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

[8] Rubén Arjona, Hai-Nan Lin, Savvas Nesseris, and Li Tang. 2021. Machine learning forecasts of the cosmic distance duality relation with strongly lensed gravitational wave events. *Phys. Rev. D*, 103(10):103513.

[9] Sriram Balasubramanian, Naman Jain, Gaurav Jindal, Abhijeet Awasthi, and Sunita Sarawagi. 2020. What's in a name? are BERT named entity representations just as good for any other name? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 205–214, Online. Association for Computational Linguistics.

[10] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org.

[11] Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

[12] Jose Luis Bernal, Licia Verde, and Adam G. Riess. 2016. The trouble with h_0. *JCAP*, 10:019.

[13] Esben Jannik Bjerrum. 2017. Smiles enumeration as data augmentation for neural network modeling of molecules. *ArXiv preprint*, abs/1703.07076.

[14] Andrea Borghesi, Federico Baldo, and Michela Milano. 2020. Improving deep learning models via constraint-based domain knowledge: a brief survey. *ArXiv preprint*, abs/2005.10691.

[15] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[16] David Camarena and Valerio Marra. 2020. A new method to build the (inverse) distance ladder. *Mon. Not. Roy. Astron. Soc.*, 495(3):2630–2644.

[17] Sitthichok Chaichulee, Chissanupong Promchai, Tanyamai Kaewkomon, Chanon Kongkamol, Thammasin Ingviya, and Pasuree Sangsupawanich. 2022. Multi-label classification of symptom terms from free-text bilingual adverse drug reaction reports using natural language processing. *PLOS ONE*, 17(8):1–22.

[18] Akshay Chaturvedi, Swarnadeep Bhar, Soumadeep Saha, Utpal Garain, and Nicholas Asher. 2024. Analyzing semantic faithfulness of language models via input intervention on question answering. *Computational Linguistics*, 50(1):119–155.

[19] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-label image recognition with graph convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5177–5186. Computer Vision Foundation / IEEE.

[20] Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

[21] Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6.

[22] Blender Online Community. 2018. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam.

[23] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20.

[24] Antonio J. Cuesta, Licia Verde, Adam Riess, and Raul Jimenez. 2015. Calibrating the cosmic distance scale ladder: the role of the sound horizon scale and the local expansion rate as distance anchors. *Mon. Not. Roy. Astron. Soc.*, 448(4):3463–3471.

[25] G Cybenko. 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2.

[26] Tirtharaj Dash, Sharad Chitlangia, Aditya Ahuja, and Ashwin Srinivasan. 2022. A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, 12(1):1040.

[27] Robin Deits. 2015. rdeits/cryptics.

[28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[30] Konstantinos F. Dialektopoulos, Purba Mukherjee, Jackson Levi Said, and Jurgen Mifsud. 2023. Neural network reconstruction of cosmology using the pantheon compilation. *Eur. Phys. J. C*, 83(10):956.

[31] Konstantinos F. Dialektopoulos, Purba Mukherjee, Jackson Levi Said, and Jurgen Mifsud. 2024. Neural network reconstruction of scalar-tensor cosmology. *Phys. Dark Univ.*, 43:101383.

[32] Michelangelo Diligenti, Marco Gori, and Claudio Saccà. 2017. Semantic-based regularization for learning and inference. *Artificial Intelligence*, 244:143–165. Combining Constraint Solving with Mining and Learning.

[33] Emily C Dimmer, Rachael P Huntley, Yasmin Alam-Faruque, Tony Sawford, Claire O'Donovan, Maria J Martin, Benoit Bely, Paul Browne, Wei Mun Chan, Ruth Eberhardt, Michael Gardner, Kati Laiho, Duncan Legge, Michele Magrane, Klemens Pichler, Diego Poggioli, Harminder Sehra, Andrea Auchincloss, Kristian Axelsen, Marie-Claude Blatter, Emmanuel Boutet, Silvia Braconi-Quintaje, Lionel Breuza, Alan Bridge, Elizabeth Coudert, Anne Estreicher, Livia Famiglietti, Serenella Ferro-Rojas, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Janet James, Silvia Jimenez, Florence Jungo, Guillaume Keller, Phillippe Lemercier, Damien Lieberherr, Patrick Masson, Madelaine Moinat, Ivo Pedruzzi, Sylvain Poux, Catherine Rivoire, Bernd Roechert, Michael Schneider, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Lydie Bougueleret, Ghislaine Argoud-Puy, Isabelle Cusin, Paula Duek-Roggli, Ioannis Xenarios, and Rolf Apweiler. 2011. The UniProt-GO annotation database in 2011. *Nucleic Acids Res*, 40(Database issue):D565–70.

[34] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[35] Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Online. Association for Computational Linguistics.

[36] Avia Efrat, Uri Shaham, Dan Kilman, and Omer Levy. 2021. Cryptonite: A cryptic crossword benchmark for extreme ambiguity in language. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4186–4192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[37] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

[38] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 20d. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *ArXiv preprint*, abs/d.

[39] Charles Elkan. 2001. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, page 973–978, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[40] Celia Escamilla-Rivera, Maryi Carvajal, Cristian Zamora, and Martin Hendry. 2022. Neural networks and standard cosmography with newly calibrated high redshift grb observations. *JCAP*, 04(04):016.

[41] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

[42] Richard Evans and Edward Grefenstette. 2018. Learning explanatory rules from noisy data. *J. Artif. Int. Res.*, 61(1):1–64.

[43] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136.

[44] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

[45] Jun-Peng Fang and Min-Ling Zhang. 2019. Partial multi-label learning via credible label elicitation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3518–3525. AAAI Press.

[46] FIDE. 2023. International chess federation handbook.

[47] Evelyn Fix and J. L. Hodges. 1989. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238–247.

[48] Manoel V. França, Gerson Zaverucha, and Artur S. D'avila Garcez. 2014. Fast relational learning using bottom clause propositionalization with artificial neural networks. *Machine Learning*, 94(1):81–104.

[49] Wendy L. Freedman and Barry F. Madore. 2023. The cepheid extragalactic distance scale: Past, present and future. *arXiv*.

[50] Wendy L. Freedman, Barry F. Madore, Taylor Hoyt, In Sung Jang, Rachael Beaton, Myung Gyoon Lee, Andrew Monson, Jill Neeley, and Jeffrey Rich. 2020. Calibration of the tip of the red giant branch (trgb). *arXiv*.

[51] Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.

[52] Luís P.F. Garcia, André C.P.L.F. de Carvalho, and Ana C. Lorena. 2016. Noise detection in the meta-learning level. *Neurocomputing*, 176:14–25. Recent Advancements in Hybrid Artificial Intelligence Systems and its Application to Real-World Problems.

[53] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1919–1925. AAAI Press.

[54] Lorenzo Giambagli, Duccio Fanelli, Guido Risaliti, and Matilde Signorini. 2023. Nonparametric analysis of the hubble diagram with neural networks. *Astron. Astrophys.*, 678:A13.

[55] Matthew L Ginsberg. 2011. Dr. fill: Crosswords and an implemented solver for singly weighted csps. *Journal of Artificial Intelligence Research*, 42:851–886.

[56] Andrés Felipe Giraldo-Forero, Jorge Alberto Jaramillo-Garzón, and César Germán Castellanos-Domínguez. 2015. Evaluation of example-based measures for multi-label classification performance. In *Bioinformatics and Biomedical Engineering*, pages 557–564, Cham. Springer International Publishing.

[57] Eleonora Giunchiglia and Thomas Lukasiewicz. 2020. Coherent hierarchical multi-label classification networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[58] Isidro Gómez-Vargas, Joshua Briones Andrade, and J. Alberto Vázquez. 2023. Neural networks optimized by genetic algorithms in cosmology. *Phys. Rev. D*, 107(4):043509.

[59] Isidro Gómez-Vargas, Ricardo Medel Esquivel, Ricardo García-Salcedo, and J. Alberto Vázquez. 2023. Neural network reconstructions for the hubble parameter, growth rate and distance modulus. *Eur. Phys. J. C*, 83(4):304.

[60] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

[61] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

[62] Yu Han, Donna M. Rizzo, John P. Hanley, Emily L. Coderre, and Patricia A. Prelock. 2022. Identifying neuroanatomical and behavioral features for autism spectrum disorder diagnosis in children using machine learning. *PLOS ONE*, 17(7):1–19.

[63] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69.

[64] Dhiraj Kumar Hazra, Subhabrata Majumdar, Supratik Pal, Sudhakar Panda, and Anjan A. Sen. 2015. Post-planck dark energy constraints. *Phys. Rev. D*, 91:083005.

[65] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

[66] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *ArXiv preprint*, abs/1606.08415.

[67] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. 2018. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10477–10486.

[68] John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

[69] John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[70] John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

[71] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *ArXiv preprint*, abs/1503.02531.

[72] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

[73] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

[74] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2410–2420, Berlin, Germany. Association for Computational Linguistics.

[75] Wei Huang, Enhong Chen, Qi Liu, Hui Xiong, Zhenya Huang, Shiwei Tong, and Dan Zhang. 2023. Hmcnet: A general approach for hierarchical multi-label classification. *IEEE Trans. Knowl. Data Eng.*, 35(9):8713–8728.

[76] Yi Huang, Buse Gilederelі, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[77] Seung-gyu Hwang, Benjamin L'Huillier, Ryan E. Keeley, M. James Jee, and Arman Shafieloo. 2023. How to use gp: effects of the mean function and hyperparameter selection on gaussian process regression. *JCAP*, 02:014.

[78] IBM. 20i. Ibm archives: Ibm 1301 disk storage unit. *ArXiv preprint*, abs/i.

[79] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 590–597. AAAI Press.

[80] ISO/IEC 7812-1:2017. 2017. Identification cards — identification of issuers — part 1: Numbering system. Standard, International Organization for Standardization, Geneva, CH.

[81] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

[82] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

[83] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

[84] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.

[85] Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. 2019. Technical report on conversational question answering. *arXiv preprint 1909.10772*.

[86] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet

Kohli, and Demis Hassabis. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596.

[87] Passent El Kafrawy, Amr Mausad, and Heba Esmail. 2015. Experimental comparison of methods for multi-label classification in different application domains. *International Journal of Computer Applications*, 114(19):1–9.

[88] Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour. 2020. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759.

[89] Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

[90] Ryan E. Keeley, Arman Shafieloo, Gong-Bo Zhao, Jose Alberto Vazquez, and Hanwool Koo. 2021. Reconstructing the universe: Testing the mutual consistency of the pantheon and sdss/eboss bao data sets with gaussian processes. *Astron. J.*, 161(3):151.

[91] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[92] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *ArXiv preprint*, abs/2304.02643.

[93] Terry Koo, Frederick Liu, and Luheng He. 2024. Automata-based constraints for language model decoding. In *First Conference on Language Modeling*.

[94] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114.

[95] Maxat Kulmanov and Robert Hoehndorf. 2022. Deepgozero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms. *Bioinformatics*, 38:i238–i245.

[96] Saurabh Kulshreshtha, Olga Kovaleva, Namrata Shivagunde, and Anna Rumshisky. 2022. Down and across: Introducing crossword-solving as a new NLP benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2648–2659, Dublin, Ireland. Association for Computational Linguistics.

[97] Ugur Kursuncu, Manas Gaur, and Amit P. Sheth. 2019. Knowledge infused learning (K-IL): towards deep incorporation of knowledge in deep learning. *ArXiv preprint*, abs/1912.00512.

[98] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4066–4076.

[99] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981.

[100] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. 1989. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems 2, [NIPS Conference, Denver, Colorado, USA, November 27-30, 1989]*, pages 396–404. Morgan Kaufmann.

[101] Tao Li and Vivek Srikumar. 2019. Augmenting neural networks with first-order logic. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 292–302, Florence, Italy. Association for Computational Linguistics.

[102] Xiaolei Li, Ryan E. Keeley, Arman Shafieloo, and Kai Liao. 2024. A model-independent method to determine h $_0$ using time-delay lensing, quasars, and type ia supernovae. *Astrophys. J.*, 960(2):103.

[103] Seppo Linnainmaa. 1976. Taylor expansion of the accumulated rounding error.

[104] Michael L. Littman, Greg A. Keim, and Noam Shazeer. 2002. A probabilistic approach to solving crossword puzzles. *Artificial Intelligence*, 134(1):23–55.

[105] Liang Liu, Li-Juan Hu, Li Tang, and Ying Wu. 2023. Constraining the spatial curvature of the local universe with deep learning. *Res. Astron. Astrophys.*, 23(12):125012.

[106] Yang Liu, Qince Li, Kuanquan Wang, Jun Liu, Runnan He, Yongfeng Yuan, and Henggui Zhang. 2021. Automatic multi-label ecg classification with category imbalance and cost-sensitive thresholding. *Biosensors*, 11(11):453.

[107] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

[108] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *ArXiv preprint*, abs/2103.14030.

[109] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. 2017. The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6231–6239.

[110] Chris Madge and Massimo Poesio. 2024. Large language models as minecraft agents. *ArXiv preprint*, abs/2402.08392.

[111] Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. 2012. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104. Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011).

[112] Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, and Marco Gori. 2020. Lyrics: A general interface layer to integrate logic inference and deep learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 283–298, Cham. Springer International Publishing.

[113] Athanasios Masouris and Jan van Gemert. 2023. End-to-end chess recognition. *Preprint*, arXiv:2310.04086.

[114] Ahmad Mehrabi. 2023. A semi-model-independent approach to describe a cosmological database. *arXiv*.

[115] Stefano Melacci, Gabriele Ciravegna, Angelo Sotgiu, Ambra Demontis, Battista Biggio, Marco Gori, and Fabio Roli. 2022. Domain knowledge alleviates adversarial attacks in multi-label classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 44(12):9944–9959.

[116] Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

[117] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *Preprint*, arXiv:2402.06196.

[118] Amina Mollaysa, Alexandros Kalousis, Eric Bruno, and Maurits Diephuis. 2019. Learning to augment with feature side-information. In *Proceedings of The Eleventh Asian Conference on Machine Learning*, volume 101 of *Proceedings of Machine Learning Research*, pages 173–187. PMLR.

[119] Purba Mukherjee, Konstantinos F. Dialektopoulos, Jackson Levi Said, and Jurgen Mifsud. 2024. Late-time transition of m_b inferred via neural networks. *arXiv*.

[120] Purba Mukherjee and Ankan Mukherjee. 2021. Assessment of the cosmic distance duality relation using gaussian process. *Mon. Not. Roy. Astron. Soc.*, 504(3):3938–3946.

[121] Nikhil Muralidhar, Mohammad Raihanul Islam, Manish Marwah, Anuj Karpatne, and Naren Ramakrishnan. 2018. Incorporating prior domain knowledge into deep neural networks. In *IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, December 10-13, 2018*, pages 36–45. IEEE.

[122] B. Novosyadlyj, O. Sergijenko, R. Durrer, and V. Pelykh. 2014. Constraining the dynamical dark energy parameters: Planck-2013 vs wmap9. *JCAP*, 05:030.

[123] Eoin Ó Colgáin and M. M. Sheikh-Jabbari. 2021. Elucidating cosmological model dependence with h_0. *Eur. Phys. J. C*, 81(10):892.

[124] Juan de Dios Rojas Olvera, Isidro Gómez-Vargas, and Jose Alberto Vázquez. 2022. Observational cosmology with artificial neural networks. *Universe*, 8(2):120.

[125] OpenAI. 2021. Chatgpt technical report.

[126] OpenAI. 2023. Gpt-4 techincal report.

[127] Lichess Org. 2013. Lichess open database.

[128] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

[129] G. Paolacci, J. Chandler, and P. G. Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419.

[130] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

[131] Rafael B. Pereira, Alexandre Plastino, Bianca Zadrozny, and Luiz H.C. Merschmann. 2018. Correlation analysis of performance measures for multi-label classification. *Information Processing And Management*, 54(3):359–369.

[132] John P. Pestian, Chris Brew, Pawel Matykiewicz, DJ Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Biological, translational, and clinical language processing*, pages 97–104, Prague, Czech Republic. Association for Computational Linguistics.

[133] Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

[134] Andrei Popescu-Belis, Àlex Atrio, Valentin Minder, Aris Xanthos, Gabriel Luthier, Simon Mattei, and Antonio Rodriguez. 2022. Constrained language models for interactive poem generation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3519–3529, Marseille, France. European Language Resources Association.

[135] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

[136] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

[137] Ryan Riegel, Alexander Gray, Francois Luus, Naweed Khan, Ndivhuwo Makondo, Ismail Yunus Akhalwaya, Haifeng Qian, Ronald Fagin, Francisco Barahona, Udit Sharma, Shajith Ikbal, Hima Karanam, Sumit Neelam, Ankita Likhyani, and Santosh Srivastava. 2020. Logical neural networks. *Preprint*, arXiv:2006.13155.

[138] Adam G. Riess and Louise Breuval. 2023. The local value of h_0. *arXiv*.

[139] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

[140] Frank Rosenblatt. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Phychological Review*, 65(6):386–408.

[141] Josh Rozner, Christopher Potts, and Kyle Mahowald. 2021. Decrypting cryptic crosswords: Semantically complex wordplay puzzles as a target for nlp. In *Advances in Neural Information Processing Systems*, volume 34, pages 11409–11421. Curran Associates, Inc.

[142] Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *ArXiv preprint*, abs/1609.04747.

[143] Andreas Ruepp, Alfred Zollner, Dieter Maier, Kaj Albermann, Jean Hani, Martin Mokrejs, Igor V. Tetko, Ulrich Güldener, Gertrud Mannhaupt, Martin Münsterkötter, and Hans-Werner Mewes. 2004. The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic acids research*, 32 18:5539–45.

[144] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning internal representations by error propagation.

[145] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323.

[146] Abdelrahman "Boda" Sadallah, Daria Kotova, and Ekaterina Kochmar. 2024. Are llms good cryptic crossword solvers? *Preprint*, arXiv:2403.12094.

[147] Soumadeep Saha, Sutanoya Chakraborty, Saptarshi Saha, and Utpal Garain. 2024. Language models are crossword solvers. *ArXiv preprint*, abs/2406.09043. Accepted to NAACL 2025 (main).

[148] Soumadeep Saha, Utpal Garain, Arijit Ukil, Arpan Pal, and Sundeep Khandelwal. 2023. Medtric : A clinically applicable metric for evaluation of multi-label computational diagnostic systems. *PLOS ONE*, 18(8):1–19.

[149] Soumadeep Saha, Utpal Garain, Arijit Ukil, Arpan Pal, and Sundeep Khandelwal. 2024. *DOST–Domain Obedient Self-supervision for Trustworthy Multi Label Classification with Noisy Labels*, pages 117–127. Springer Nature Switzerland, Cham.

[150] Soumadeep Saha, Saptarshi Saha, and Utpal Garain. 2024. VALUED - vision and logical understanding evaluation dataset. *Journal of Data-centric Machine Learning Research*.

[151] Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):1–21.

[152] Terence Sanger and Pallavi N. Baljekar. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408.

[153] Robert E. Schapire and Yoram Singer. 2000. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39:135–168.

[154] Hendrik Schuff, Heike Adel, and Ngoc Thang Vu. 2020. F1 is Not Enough! Models and Evaluation Towards User-Centered Explainable Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7076–7095, Online. Association for Computational Linguistics.

[155] Bernhard Schölkopf. 2019. Causality for machine learning. *ArXiv preprint*, abs/1911.10500.

[156] D. M. Scolnic, D. O. Jones, A. Rest, Y. C. Pan, R. Chornock, R. J. Foley, M. E. Huber, R. Kessler, G. Narayan, A. G. Riess, S. Rodney, E. Berger, D. J. Brout, P. J. Challis, M. Drout, D. Finkbeiner, R. Lunnan, R. P. Kirshner, N. E. Sanders, E. Schlafly, S. Smartt, C. W. Stubbs, J. Tonry, W. M. Wood-Vasey, M. Foley, J. Hand, E. Johnson, W. S. Burgett, K. C. Chambers, P. W. Draper, K. W. Hodapp, N. Kaiser, R. P. Kudritzki, E. A. Magnier, N. Metcalfe, F. Bresolin, E. Gall, R. Kotak, M. McCrum, and K. W. Smith. 2020. The Complete Light-curve Sample of Spectroscopically Confirmed SNe Ia from Pan-STARRS1 and Cosmological Constraints from the Combined Pantheon Sample. *ArXiv preprint*, abs/2018a.

[157] Dan Scolnic, Dillon Brout, Anthony Carr, Adam G. Riess, Tamara M. Davis, Arianna Dwomoh, David O. Jones, Noor Ali, Pranav Charvu, Rebecca Chen, Erik R. Peterson, Brodie Popovic, Benjamin M. Rose, Charlotte M. Wood, Peter J. Brown, Ken Chambers, David A. Coulter, Kyle G. Dettman, Georgios Dimitriadis, Alexei V. Filippenko, Ryan J. Foley, Saurabh W. Jha, Charles D. Kilpatrick, Robert P. Kirshner, Yen-Chen Pan, Armin Rest, Cesar Rojas-Bravo, Matthew R. Siebert, Benjamin E. Stahl, and WeiKang Zheng. 2022. The pantheon+ analysis: The full data set and light-curve release. *The Astrophysical Journal*, 938(2):113.

[158] T. Segaran and J. Hammerbacher. 2009. *Beautiful Data: The Stories Behind Elegant Data Solutions*. Theory in practice. O'Reilly Media.

[159] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

[160] Luciano Serafini and Artur S. d'Avila Garcez. 2016. Logic tensor networks: Deep learning and logical reasoning from data and knowledge. In *Proceedings of the 11th International Workshop on Neural-Symbolic Learning and Reasoning (NeSy'16)*, volume 1768 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[161] Rahul Shah, Purba Mukherjee, Soumadeep Saha, Utpal Garain, and Supratik Pal. 2024. Deep learning based recalibration of sdss and desi bao alleviates hubble and clustering tensions. *ArXiv preprint*, abs/2412.14750.

[162] Rahul Shah, Soumadeep Saha, Purba Mukherjee, Utpal Garain, and Supratik Pal. 2024. Ladder: Revisiting the cosmic distance ladder with deep learning approaches and exploring its applications. *The Astrophysical Journal Supplement Series*, 273(2):27.

[163] Torgyn Shaikhina and Natalia A. Khovanova. 2017. Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial Intelligence in Medicine*, 75:51–63.

[164] Noam M. Shazeer, Michael L. Littman, and Greg A. Keim. 1999. Solving crossword puzzles as probabilistic constraint satisfaction. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, AAAI '99/IAAI '99, page 156–162, USA. American Association for Artificial Intelligence.

[165] Ryan Sheatsley, Blaine Hoak, Eric Pauley, Yohan Beugin, Michael J. Weisman, and Patrick McDaniel. 2021. On the robustness of domain constraints. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, page 495–515, New York, NY, USA. Association for Computing Machinery.

[166] Mattia Silvestri, Michele Lombardi, and Michela Milano. 2021. Injecting domain knowledge in neural networks: A controlled experiment on a constrained problem. In *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pages 266–282, Cham. Springer International Publishing.

[167] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[168] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–19.

[169] Liwei Song, Xinwei Yu, Hsuan-Tung Peng, and Karthik Narasimhan. 2021. Universal adversarial attacks with natural triggers for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3724–3733, Online. Association for Computational Linguistics.

[170] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

[171] Rick Stevens, Valerie Taylor, Jeff Nichols, Arthur Barney Maccabe, Katherine Yelick, and David Brown. 2020. AI for science: Report on the department of energy (DOE) town halls on artificial intelligence (AI) for science.

[172] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C).

[173] Lichao Sun, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip S. Yu, and Caiming Xiong. 20 2. Adv-bert: BERT is not robust on misspellings! generating nature adversarial samples on BERT. *ArXiv preprint*, abs/ 2003.

[174] Lijuan Sun, Songhe Feng, Tao Wang, Congyan Lang, and Yi Jin. 2019. Partial multi-label learning by low-rank and sparse decomposition. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 5016–5023. AAAI Press.

[175] Mirac Suzgun, Luke Melas-Kyriazi, Suproteem K. Sarkar, Scott Duke Kominers, and Stuart M. Shieber. 2022. The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications. *ArXiv preprint*, abs/2207.04043.

[176] Li Tang, Xin Li, Hai-Nan Lin, and Liang Liu. 2021. Model-independently calibrating the luminosity correlations of gamma-ray bursts using deep learning. *Astrophys. J.*, 907(2):121.

[177] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, and 1369 Others. 2023. Gemini: A family of highly capable multimodal models. *ArXiv preprint*, abs/2312.11805.

[178] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

[179] Nguyen Thai-Nghe, Zeno Gantner, and Lars Schmidt-Thieme. 2010. Cost-sensitive learning methods for imbalanced data. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

[180] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

[181] Geoffrey G. Towell and Jude W. Shavlik. 1994. Knowledge-based artificial neural networks. *Artificial Intelligence*, 70(1):119–165.

[182] Arash Vahdat. 2017. Toward robustness against label noise in training deep discriminative neural networks. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5596–5605.

[183] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

[184] Matt Visser. 2005. Cosmography: Cosmology without the einstein equations. *Gen. Rel. Grav.*, 37:1541–1548.

[185] Khuong Vo, Tri Nguyen, Dang Pham, Mao Nguyen, Minh Truong, Trung Mai, and Tho T. Quan. 2010. Combination of domain knowledge and deep learning for sentiment analysis of short and informal messages on social media. *ArXiv preprint*, abs/10.1504.

[186] Phong D. Vo, Alexandru-Lucian Gînsca, Hervé Le Borgne, and Adrian Daniel Popescu. 2015. Effective training of convolutional networks using noisy web images. *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6.

[187] Eric Wallace, Nicholas Tomlin, Albert Xu, Kevin Yang, Eshaan Pathak, Matthew Ginsberg, and Dan Klein. 2022. Automated crossword solving. In *ACL:2022:long*, pages 3073–3085, Dublin, Ireland. acl.

[188] Ge Wang, Jong Chu Ye, Klaus Mueller, and Jeffrey A Fessler. 2018. Image reconstruction is a new frontier of machine learning. *IEEE Trans Med Imaging*, 37(6):1289–1296.

[189] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*.

[190] Guo-Jian Wang, Si-Yao Li, and Jun-Qing Xia. 2020. Ecopann: A framework for estimating cosmological parameters using artificial neural networks. *Astrophys. J. Suppl.*, 249(2):25.

[191] Guo-Jian Wang, Xiao-Jiao Ma, Si-Yao Li, and Jun-Qing Xia. 2020. Reconstructing functions and estimating parameters with artificial neural networks: A test with a hubble parameter and sne ia. *Astrophys. J. Suppl.*, 246(1):13.

[192] Huazhen Wang, Xin Liu, Bing Lv, Fan Yang, and Yanzhu Hong. 2014. Reliable multi-label learning via conformal predictor and random forest for syndrome differentiation of chronic fatigue in traditional chinese medicine. *PLOS ONE*, 9(6):1–14.

[193] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

[194] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. 2020. Multi-label classification with label graph superimposing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12265–12272.

[195] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

[196] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

[197] 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.

[198] Ronald J. Williams and David Zipser. 1995. Gradient-based learning algorithms for recurrent networks and their computational complexity.

[199] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv preprint*, abs/1910.03771.

[200] Georg Wölflein and Ognjen Arandjelović. 2021. Determining chess game state from an image. *Journal of Imaging*, 7(6):94.

[201] Hanbei Xie, Xiaodong Nong, Bin Zhang, Huifeng Wang, Zihao Li, and Nan Liang. 2023. Constraints on cosmological models with gamma-ray bursts in cosmology-independent way. *arXiv*.

[202] Ming-Kun Xie and Sheng-Jun Huang. 2018. Partial multi-label learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4302–4309. AAAI Press.

[203] Ming-Kun Xie and Sheng-Jun Huang. 2022. Partial multi-label learning with noisy label identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3676–3687.

[204] Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Zhengsu Chen, Shaojie Tang, and Shui Yu. 2021. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, 69:101985.

[205] Yaqi Xie, Ziwei Xu, Kuldeep S. Meel, Mohan S. Kankanhalli, and Harold Soh. 2019. Embedding symbolic knowledge into deep networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information*

Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 4235–4245.

[206] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van den Broeck. 2018. A semantic loss function for deep learning with symbolic knowledge. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5498–5507. PMLR.

[207] Shweta Yadav, Usha Lokala, Raminta Daniulaityte, Krishnaprasad Thirunarayan, Francois Lamy, and Amit Sheth. 2021. "when they say weed causes depression, but it's your fav antidepressant": Knowledge-aware attention framework for relationship extraction. *PLOS ONE*, 16(3):1–18.

[208] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

[209] 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, Brussels, Belgium. Association for Computational Linguistics.

[210] Zijiang Yang, Reda Al-Bahrani, Andrew C. E. Reid, Stefanos Papanikolaou, Surya R. Kalidindi, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. 2019. Deep learning based domain knowledge integration for small datasets: Illustrative applications in materials informatics. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. Material Science.

[211] Yazhou Yao, Fumin Shen, Guosen Xie, Li Liu, Fan Zhu, Jian Zhang, and Heng Tao Shen. 2020. Exploiting web images for multi-output classification: From category to subcategories. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2348–2360.

[212] Bin Zhang, Xiaoyao Xie, Xiaodong Nong, Guangzhen Wang, Zhiguo Xiong, Puxun Wu, and Nan Liang. 2023. Model-independent gamma-ray bursts constraints on cosmological models using machine learning. *arXiv*.

[213] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. RAVEN: A dataset for relational and analogical visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5317–5327. Computer Vision Foundation / IEEE.

[214] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115.

[215] Jian-Chen Zhang, Yu Hu, Kang Jiao, Hong-Feng Wang, Yuan-Bo Xie, Bo Yu, Li-Li Zhao, and Tong-Jie Zhang. 2024. A nonparametric reconstruction of the hubble parameter h(z) based on radial basis function neural networks. *Astrophys. J. Suppl.*, 270(2):23.

[216] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

[217] Yinyuan Zhang, Ricardo Henao, Zhe Gan, Yitong Li, and Lawrence Carin. 2018. Multi-label learning from medical plain text with convolutional residual models. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 280–294. PMLR.

[218] Wenting Zhao and Carla P. Gomes. 2021. Evaluating multi-label classifiers with noisy labels. *ArXiv preprint*, abs/2102.08427.

[219] Liang Zhou, Xiaoyuan Zheng, Di Yang, Ying Wang, Xuesong Bai, and Xinhua Ye. 2021. Application of multi-label classification models for the diagnosis of diabetic complications. *BMC Medical Informatics and Decision Making*, 21(1):182.

[220] Ke Zhu and Jianxin Wu. 2021. Residual attention: A simple but effective method for multi-label recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 184–193.

[221] Shuang-Li Zhu, Jie Dong, Chenjing Zhang, Yao-Bo Huang, and Wensheng Pan. 2021. Application of machine learning in the diagnosis of gastric cancer based on noninvasive characteristics. *PLOS ONE*, 15(12):1–13.

[222] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2019. A comprehensive survey on transfer learning. *ArXiv preprint*, abs/1911.02685.