Scalable Understanding of Datasets and Models with the Help of Language Models

Ruiqi Zhong*

May 9, 2025

Preface

This document provides a list of useful citations related to my tutorial. For a self-contained written introduction, please refer to my Ph.D. thesis: *Natural Language Explanations of Dataset Patterns*.

Using language models to explain other models and datasets is an important and promising research direction. I hope future researchers will build upon the works mentioned in this document. As this is currently a medium-sized research field, I cannot cover all related papers here, and this document should not be considered a comprehensive survey. See Singh et al. (2024) for more related works.

0 Intro

Predictive Power of Explanations. The core idea we introduced in this line of research is that "good explanations should allow humans to predict empirical observations." This idea is well-established in scientific instrumentalism[1], where scientific explanations do not aim to discover what is "inherently true," but rather to help humans predict empirical phenomena and accomplish useful tasks.

Limitations of Traditional Explanation Methods. To categorize content in a large text corpus, topic models such as LDA (Blei et al., 2003) represent each topic as weights over a high-dimensional word vector, but Chang et al. (2009) shows that the learned topics sometimes lack coherent meaning. Our work (Wang et al., 2023) corroborated this finding and demonstrated that cluster explanations based on natural language are inherently more interpretable to humans.

To explain image classifiers, gradient-based interpretability methods such as Grad-Cam (Selvaraju et al., 2017) highlight sets of pixels to explain classifier decisions. However, many concepts cannot be effectively explained by merely highlighting certain subregions.

Applications of Using LLMs to Explain. There are already many real-world applications of using language models for explanation. We cover several mentioned in our tutorial:

*Work done at UC Berkeley as a Ph.D. student, current email: ruiqizhong1997@gmail.com.

[1] `https://en.wikipedia.org/wiki/Instrumentalism`

- **Explaining LLM Neurons:** Choi et al. (2024); Bills et al. (2023) use language models to explain language model neurons, and Meng et al. (2024) demonstrates how neuron descriptions can help debug LLM behaviors.
- **Understanding Real-World AI Use**: Tamkin et al. (2024) applies explainable clustering to user conversations with chatbots to understand real-world use cases, and follow-up research Handa et al. (2025) uses such a system to understand the economic value of chatbots.
- **Understanding LLM Output Style**: Dunlap et al. (2025); Sun et al. (2025) compare the output styles (tone, formatting, and writing style) of different language models, aspects not covered by traditional evaluations.

1 Explaining Static Datasets

1.1 Explaining Dataset Differences

Core Method: Proposer-Validator Framework. The canonical approach to explaining datasets is to first propose explanations with LLMs and then validate them on individual datapoints. This approach is useful not only for explaining dataset differences but also for "guessing the underlying instruction." Multiple papers have independently discovered and demonstrated the utility of this approach, including Singh et al. (2022); Zhou et al. (2022); Zhong et al. (2023, 2022); Honovich et al. (2022).

Multi-Modal Application. The proposer-validator approach works not only in the text domain but also in the image domain. Dunlap et al. (2023) uses this approach to describe differences between sets of images, and Zhu et al. (2022) specifically focuses on explaining distribution shifts in image datasets.

Extension 1: Precise explanations. In Wang et al. (2023); Zhong et al. (2024), we generate more detailed explanations by modifying the proposer prompt.

Extension 2: Goal-Constrained Explanations. In Zhong et al. (2023); Wang et al. (2023), we propose adding natural language constraints to the prompt so that the explanations can be useful for the goal.

Extension 3: Extracting Multiple Explanations. In Zhong et al. (2024), we propose extracting multiple explanations by fitting a linear model on top of $[\![\phi]\!](X)$. An independent line of work on "language-based concept bottleneck networks" has also explored similar approaches (Yang et al., 2023; Ludan et al., 2023; Chiquier et al., 2024; Schrodi et al., 2024).

1.2 Explainable Clustering

Many papers have explored using LLMs for clustering, including Wang et al. (2023); Pham et al. (2023); Lam et al. (2024); Viswanathan et al. (2024); Zhong et al. (2024). The propose and validation (assignment) steps appear in all of these papers.

2 Explaining Models

We touched on four different kinds of model explanations: 1) explaining neurons, 2) categorizing output styles, 3) explaining decision boundaries, and 4) understanding what LLMs are good at. We have covered related works for 1) and 2) in Section 0, and we will cover 3) and 4) here.

Explaining Decision Boundaries. Our presentation is mostly based on the content from Chen et al. (2024b), which uses counterfactual simulatability to evaluate model explanations. Mills et al. (2023) explored a similar approach.

Understanding What LLMs Are Good At. Given a dataset of inputs and how a target LM performs on these inputs, Sobhani et al. (2025); Zhong et al. (2024); Zeng et al. (2025) use LM-based systems to explain categories of inputs where the target LM underperforms.

3 Future Directions

Validation Efficiency. Currently, validation is inefficient since we need to call a language model to validate each pair of ϕ and x. It might be possible to borrow ideas from the retrieval literature Khattab and Zaharia (2020), where they first embed ϕ and x independently and then perform lightweight pair-wise computation on them.

Proposer Efficiency. We can use the validation score as the reward to train the proposer so that they can directly output better explanations. Choi et al. (2024) has explored this approach to better describe neurons, and Chen et al. (2024a) has explored this to propose better self-explanations.

Concepts that Do Not Have Words for Yet. Hewitt et al. (2025) explored this issue in greater details.

References

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., and Saunders, W. (2023). Language models can explain neurons in language models. *URL https://openaipublic. blob. core. windows. net/neuron-explainer/paper/index. html.(Date accessed: 14.05. 2023)*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

Chen, Y., Singh, C., Liu, X., Zuo, S., Yu, B., He, H., and Gao, J. (2024a). Towards consistent natural-language explanations via explanation-consistency finetuning. *arXiv preprint arXiv:2401.13986*.

Chen, Y., Zhong, R., Ri, N., Zhao, C., He, H., Steinhardt, J., Yu, Z., and Mckeown, K. (2024b). Do models explain themselves? Counterfactual simulatability of natural language explanations. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 7880–7904. PMLR.

Chiquier, M., Mall, U., and Vondrick, C. (2024). Evolving interpretable visual classifiers with large language models. *arXiv preprint arXiv:2404.09941*.

Choi, D., Huang, V., Meng, K., Johnson, D. D., Steinhardt, J., and Schwettmann, S. (2024). Scaling automatic neuron description. `https://transluce.org/neuron-descriptions`.

Dunlap, L., Mandal, K., Darrell, T., Steinhardt, J., and Gonzalez, J. E. (2025). Vibecheck: Discover and quantify qualitative differences in large language models. In *The Thirteenth International Conference on Learning Representations*.

Dunlap, L., Zhang, Y., Wang, X., Zhong, R., Darrell, T., Steinhardt, J., Gonzalez, J. E., and Yeung-Levy, S. (2023). Describing differences in image sets with natural language. *arXiv preprint arXiv:2312.02974*.

Handa, K., Tamkin, A., McCain, M., Huang, S., Durmus, E., Heck, S., Mueller, J., Hong, J., Ritchie, S., Belonax, T., et al. (2025). Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761*.

Hewitt, J., Geirhos, R., and Kim, B. (2025). We can't understand ai using our existing vocabulary. *arXiv preprint arXiv:2502.07586*.

Honovich, O., Shaham, U., Bowman, S. R., and Levy, O. (2022). Instruction induction: From few examples to natural language task descriptions. *arXiv preprint arXiv:2205.10782*.

Khattab, O. and Zaharia, M. (2020). Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Lam, M. S., Teoh, J., Landay, J. A., Heer, J., and Bernstein, M. S. (2024). Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–28.

Ludan, J. M., Lyu, Q., Yang, Y., Dugan, L., Yatskar, M., and Callison-Burch, C. (2023). Interpretable-by-design text classification with iteratively generated concept bottleneck. *arXiv preprint arXiv:2310.19660*.

Meng, K., Huang, V., Chowdhury, N., Choi, D., Steinhardt, J., and Schwettmann, S. (2024). Monitor: An ai-driven observability interface. `https://transluce.org/observability-interface`. Technical demonstration.

Mills, E., Su, S., Russell, S., and Emmons, S. (2023). Almanacs: A simulatability benchmark for language model explainability. *arXiv preprint arXiv:2312.12747*.

Pham, C. M., Hoyle, A., Sun, S., and Iyyer, M. (2023). Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*.

Schrodi, S., Schur, J., Argus, M., and Brox, T. (2024). Concept bottleneck models without predefined concepts. *arXiv preprint arXiv:2407.03921*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Singh, C., Inala, J. P., Galley, M., Caruana, R., and Gao, J. (2024). Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.

Singh, C., Morris, J. X., Aneja, J., Rush, A. M., and Gao, J. (2022). Explaining patterns in data with language models via interpretable autoprompting. *arXiv preprint arXiv:2210.01848*.

Sobhani, D., Zhong, R., Marrese-Taylor, E., Sakaguchi, K., and Matsuo, Y. (2025). Language models can categorize system inputs for performance analysis. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6241–6257, Albuquerque, New Mexico. Association for Computational Linguistics.

Sun, M., Yin, Y., Xu, Z., Kolter, J. Z., and Liu, Z. (2025). Idiosyncrasies in large language models. *arXiv preprint arXiv:2502.12150*.

Tamkin, A., McCain, M., Handa, K., Durmus, E., Lovitt, L., Rathi, A., Huang, S., Mountfield, A., Hong, J., Ritchie, S., et al. (2024). Clio: Privacy-preserving insights into real-world ai use. *arXiv preprint arXiv:2412.13678*.

Viswanathan, V., Gashteovski, K., Gashteovski, K., Lawrence, C., Wu, T., and Neubig, G. (2024). Large language models enable few-shot clustering. *Transactions of the Association for Computational Linguistics*, 12:321–333.

Wang, Z., Shang, J., and Zhong, R. (2023). Goal-driven explainable clustering via language descriptions. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10626–10649, Singapore. Association for Computational Linguistics.

Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., and Yatskar, M. (2023). Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197.

Zeng, Z., Wang, Y., Hajishirzi, H., and Koh, P. W. (2025). Evaltree: Profiling language model weaknesses via hierarchical capability trees. *arXiv preprint arXiv:2503.08893*.

Zhong, R., Snell, C., Klein, D., and Steinhardt, J. (2022). Describing differences between text distributions with natural language. In *International Conference on Machine Learning*, pages 27099–27116. PMLR.

Zhong, R., Wang, H., Klein, D., and Steinhardt, J. (2024). Explaining datasets in words: Statistical models with natural language parameters. *Advances in Neural Information Processing Systems*, 37:79350–79380.

Zhong, R., Zhang, P., Li, S., Ahn, J., Klein, D., and Steinhardt, J. (2023). Goal driven discovery of distributional differences via language descriptions. *arXiv preprint arXiv:2302.14233*.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. (2022). Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

Zhu, Z., Liang, W., and Zou, J. (2022). Gsclip: A framework for explaining distribution shifts in natural language. *arXiv preprint arXiv:2206.15007*.