Shiyuan Huang

University of California, Santa Cruz
1156 High St
Santa Cruz
California, United States, 95064

shuan101@ucsc.edu
https://shiyuan-eric.github.io/website/

1 Research interests

The rapid growth in the development of generative AI models has made their evaluation as crucial as uncovering their generative capabilities, such as audio text, audio, image and video generation. My research is focused on analyzing these models in terms of their **explainability**, **interpretability**, and **trustworthiness**.

Explainability focuses on the decision-making processes of these models. My research seeks to answer the question: Can the model explain how it made a particular decision? Additionally, it explores what can help the model generate meaningful and understandable explanations about the reasons behind its predictions. Given the nature of neural networks, analyzing parameters in each neuron is often unproductive. Therefore, various methods, such as post-hoc analysis, have been developed to address this question from different angles. However, many methods, such as post-hoc analysis, merely scratch the surface of neural networks. Much further research is needed to address the numerous unresolved problems in this emerging field.

Interpretability involves understanding the inner workings of the models. Given their powerful generative capabilities, it is challenging to determine whether the model has fully comprehended all requirements and generated accurate content, especially when the user is unsure of the correct answer. Thus, I am interested in causal tracing , such as mechanistic interpretability, to gain a deeper understanding of the models.

Both explainability and interpretability aim to achieve the same goal: understanding the generation process and explaining the capabilities of generative models. This understanding will enhance user experience by increasing trust in and effective utilization of the models' outputs, which leads to the aspect of **trustworthiness**.

Given the discussion of research concepts that I am interested in, here are some methods and applications that utilize these concepts:

1.1 Traditional AI Methods + Generative AI

With the long time development of AI, there are many methods being well studied in the field of explainability. For instance, feature attribution is a method that being used to determine how much each feature in a model contribute to the evaluation. My prior work (Huang et al., 2023) tests the performance of language models by using feature attribution method. And it turns out that many traditional AI methods are not well-suited to super-power generative AI. For instance, as a human, sometimes what matters most in a classification problem is not a specific feature but the overall impression from many features. It will be a very potential topic to adapt traditional AI methods to new generative AI model and improve them to better fit this type of high intelligent models.

1.2 Generative AI + Logic

Generative AI is powerful yet unpredictable. Meanwhile, logic is good at eliminating uncertainty. There are many existing works in the field of Neural Symbolic AI that tried to combine logic to deep neural network (Yang et al., 2023; Zhang et al., 2023). Especially, there are many works that use logic to improve the explainability of models such as adding a external knowledge base in the model so that every part of the output can be traced from the knowledge bases (Razniewski et al., 2021; Sun et al., 2021). Slightly different than this, I am interested in trying new methods to incorporate logic with generative AI to improve the explainability and interpretability. For instance, could we use logic as a helpful guide, without strictly adhering to its rules, to enhance the ability of language models to generate better explanations?

1.3 Education Application

I am also interested in the application of generative AI in the field of education (Niousha et al., 2024). There has been a trend that students tend to use generative AI like ChatGPT to seeking for answers instead of searching for answers online. From my point of view, the difference between asking ChatGPT and search engine is very similar to learning via teacher and textbook. The users tend to prefer a more structure, with just right amount of information instead of being overwhelmed. I am interested in applying the power of generative AI to be a personal tutor that not only provides correct solutions but also provides customized feedback for different users.

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 62–63

2 Spoken dialogue system (SDS) research

I will never underestimate the rapid development of leading research in dialogue systems. My prediction for dialogue research in the next 5 to 10 years is that we will expand beyond human language, potentially being applied in the field of biology, particularly, studying the dialog system of animals like whales, birds, and dogs. With this field of research being explored, there will be chance that in the future, there will be no language barrier between animals and humans.

Based on the current trajectory of SDS research, I have observed an increasing focus on developing systems with high accuracy and effective penalization mechanisms. With the success of language models, I am confident that SDS will emerge as a significant trend in the coming years. Language models offer a robust foundation by generating the necessary content, but the next challenge lies in ensuring proper verbal delivery. For example, how should an SDS respond when a user interrupts it mid-response? Should it continue or start over? An SDS must not only deliver accurate answers but also communicate in a natural manner.

3 Suggested topics for discussion

- With the advancement of generative AI models, their capabilities have expanded significantly. These models now support inputs and outputs across various modalities, including text, voice, image, and video. However, a major challenge in this field is the lack of a definitive ground truth for evaluation. Traditional AI models often rely on a single correct answer to assess performance. In contrast, generative models produce a range of possible outputs, making it nearly impossible to pinpoint just one correct solution for a given task. Therefore, developing innovative evaluation metrics, particularly in the field of SDS, will be a crucial area of research.

- Generative AI has shown some impressive abilities in many areas. But often, its full potential is not being completely utilized or it's being used in ways that don't quite fit. For instance, in the field of education, people tend to use generative AI as a knowledgeable search engine, which does not really take advantage of what it can do. The same goes for SDS, where figuring out the best way to unlock all their power is a challenge worth discussing.

- Another pressing concern is the wrongful usage associated with dialogue systems that include voice recreation which can be exploited for scams by mimicking someone's voice or using one singer's voice for another song. Looking ahead, regulations need to be refined to better handle problems such as the misuse of content creation that infringes on copyrights or is used in scams. Exploring how these regulations can be improved to address such issues will be crucial for the responsible development and deployment of dialogue systems.

References

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207* .

Rose Niousha, Muntasir Hoq, Bita Akram, and Narges Norouzi. 2024. Use of large language models for extracting knowledge components in cs1 programming exercises. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*. pages 1762–1763.

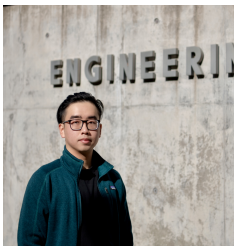
Simon Razniewski, Andrew Yates, Nora Kassner, and Gerhard Weikum. 2021. Language models as or for knowledge bases. *arXiv preprint arXiv:2110.04888* .

Haitian Sun, Patrick Verga, Bhuwan Dhingra, Ruslan Salakhutdinov, and William W Cohen. 2021. Reasoning over virtual knowledge bases with open predicate relations. In *International Conference on Machine Learning*. PMLR, pages 9966–9977.

Zhun Yang, Adam Ishay, and Joohyung Lee. 2023. Coupling large language models with logic programming for robust and general reasoning from text. *arXiv preprint arXiv:2307.07696* .

Hanlin Zhang, Jiani Huang, Ziyang Li, Mayur Naik, and Eric Xing. 2023. Improved logical reasoning of language models via differentiable symbolic programming. *arXiv preprint arXiv:2305.03742* .

Biographical sketch



Shiyuan Huang is a first-year PhD student at the University of California, Santa Cruz. His research focuses on the explainability and interpretability of language models. Prior to his PhD studies, Shiyuan completed his master's degree at the University of California, Santa Cruz. His master's research project, titled "Can Large Language Models Explain Themselves? A Study of LLM-Generated Self-Explanations," explored the use of feature attribution methods with large language models to measure their explainability.