Explanation in the Era of LLMs

NAACL 2024 tutorial
Website: https://explanation-llm.github.io/
Conclusion and Discussion

Outline of the tutorial

1. Motivation and desiderata
2. Prompting-based Explanations
3. Data attribution
4. Transformer understanding
5. **Conclusion and discussion**

Explanation methods to address the challenges

	Prompt-based	Influence function	Mech interp
Lack of transparency	Improve communication, faithfulness, etc	Understand data	Many ways
Computation	Language understanding and instruction following	Understand and refine data	Understand model
Unstability	Improve reasoning, grounding and instructability	Improve stability	Facilitate targeted control
Over-reliance	Improve communication, faithfulness, etc	Help understand the supports for model decisions	Help understand mechanisms of model decisions
Hard to evaluate	New evaluation methods and paradigms	New evaluation methods and paradigms	New evaluation methods and paradigms

Explanation methods to address the challenges

	Prompt-based	Influence function	Mech interp
Lack of transparency	Improve communication, faithfulness, etc	Understand data	Many ways
Computation	Language understanding and instruction following	Understand and refine data	Understand model
Unstability	Improve reasoning, grounding and instructability	Improve stability	Facilitate targeted control
Over-reliance	Improve communication, faithfulness, etc	Help understand the supports for model decisions	Help understand mechanisms of model decisions
Hard to evaluate	New evaluation methods and paradigms	New evaluation methods and paradigms	New evaluation methods and paradigms

Explanation methods to address the challenges

	Prompt-based	Influence function	Mech interp
Lack of transparency	Improve communication, faithfulness, etc	Understand data	Many ways
Computation	Language understanding and instruction following	Understand and refine data	Understand model
Unstability	Improve reasoning, grounding and instructability	Improve stability	Facilitate targeted control
Over-reliance	Improve communication, faithfulness, etc	Help understand the supports for model decisions	Help understand mechanisms of model decisions
Hard to evaluate	New evaluation methods and paradigms	New evaluation methods and paradigms	New evaluation methods and paradigms

Explanation methods to address the challenges

	Prompt-based	Influence function	Mech interp
Lack of transparency	Improve communication, faithfulness, etc	Understand data	Many ways
Computation	Language understanding and instruction following	Understand and refine data	Understand model
Unstability	Improve reasoning, grounding and instructability	Improve stability	Facilitate targeted control
Over-reliance	Improve communication, faithfulness, etc	Help understand the supports for model decisions	Help understand mechanisms of model decisions
Hard to evaluate	New evaluation methods and paradigms	New evaluation methods and paradigms	New evaluation methods and paradigms

Explanation methods to address the challenges

	Prompt-based	Influence function	Mech interp
Lack of transparency	Improve communication, faithfulness, etc	Understand data	Many ways
Computation	Language understanding and instruction following	Understand and refine data	Understand model
Unstability	Improve reasoning, grounding and instructability	Improve stability	Facilitate targeted control
Over-reliance	Improve communication, faithfulness, etc	Help understand the supports for model decisions	Help understand mechanisms of model decisions
Hard to evaluate	New evaluation methods and paradigms	New evaluation methods and paradigms	New evaluation methods and paradigms

Final Q&A: Explanation in the Era of LLM

Tutorial Website: https://explanation-llm.github.io/

Zining Zhu
SIT

Hanjie Chen
Rice

Xi Ye
UT Austin

Chenhao Tan
UChicago

Ana Marasović
Utah

Sarah Wiegreffe
AI2

Veronica Qing Lyu
UPenn