Explanation in the Era of LLMs

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

Presenters

Zining Zhu
SIT

Hanjie Chen
Rice

Xi Ye
UT Austin

Chenhao Tan
UChicago

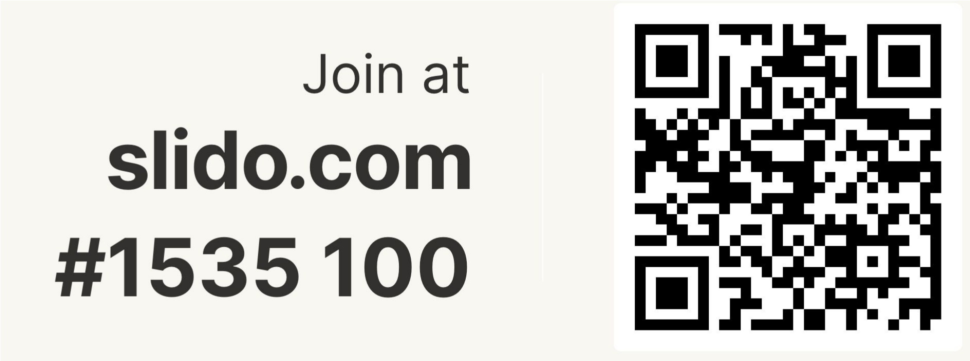
Ana Marasović
Utah

Sarah Wiegreffe
AI2

Veronica Qing Lyu
UPenn

Live Q&A

We will collect questions here.



Join at

slido.com

#1535 100

Why is that a solution to this problem?

Q: A baseball player throws a ball straight upwards at a speed of 15 m/s. Ignoring air resistance, how high will the ball go before it starts coming back down?

A: 11.48m

Throughout the process that the ball goes up, the total energy remains constant, following the conservation of energy.

At the highest point, the initial kinetic energy is all converted to the gravitational potential energy: $(1/2)mv^2 = mgh$

So the maximum height is $h=v^2/(2g)=11.48$m

Why did you do that?

A CEO, Alice, is swamped preparing for a pitch to the investors. There's an upcoming product launch event from a competitor.

The assistant, Bob, finds a research analyst to attend that event, and takes off this product launch event from Alice's calendar.

Here's Bob's explanation for this behavior:

I took the liberty of booking our research analyst, Maria, to attend. Having her there could give us valuable insights into their new product's features and potential weaknesses. This information could be crucial for tailoring your pitch to highlight our competitive advantage.

What is the stuff that follows the “why”?

Let’s clarify them in this example:

Why is the sky turning dark? The sun is going down.

- **Explanandum** is what describes the **problem**.
- **Explanan** is what describes the **reason**.
- **Explanation** is the **process(*)** of stating the reason for the problem.

*: The term “**explanation**” is also frequently used to refer to the **product** of this process.

Explanation tells the reason

Explanation has been associated with **causation**.

“The explanandum must be a **logical consequence** of the explanans.” [Hempel & Oppenheim, 1948]

“Not all explanations are about causal relations [...] Yet the vast majority of our everyday explanations invoke notions of **cause and effect**.” [Keil, 2006]

Carl G. Hempel and Paul Oppenheim. 1948. Studies in the Logic of Explanation. *Philosophy of Science*, 15(2):135–175.

Frank C. Keil. 2006. Explanation and Understanding. *Annual review of psychology*, 57:227–254.

Explanation is a process involving humans

“Explanations are often between individuals and reflect an attempt to **communicate an understanding**.” [Keil, 2006]

“Explanations are social – they are **a transfer of knowledge**, presented as part of a conversation or interaction, and are thus presented relative to the explainer's beliefs about the explainee's beliefs.” [Miller, 2018]

Frank C. Keil. 2006. Explanation and Understanding. *Annual review of psychology*, 57:227–254.

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

There are many types of explanations

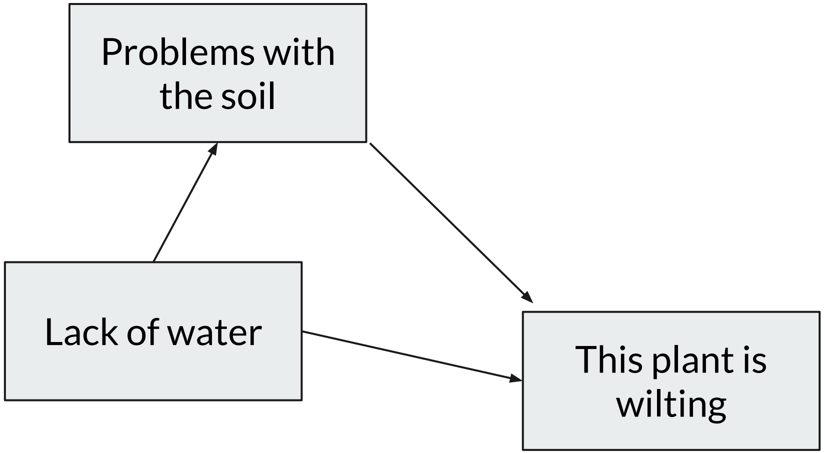
By the **reasoning patterns**:

A plant in the living room is wilting. Why?

- **Deductive**: If a plant does not receive enough water, it will wilt. That plant in the living room hasn’t be watered for a month.
- **Inductive**: Other plants in that living room have wilted because they haven’t received enough water. It’s likely that plants wilt because of lack of water.
- **Abductive**: It’s likely that the plant in this living room is wilting because there’s no enough water.
- …

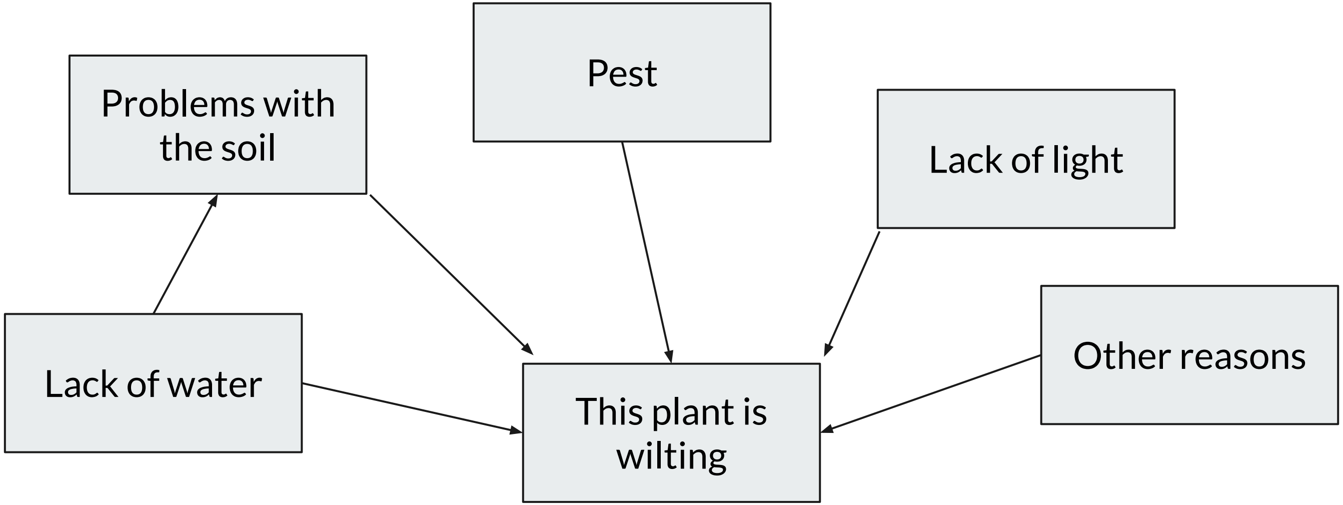
Many types of explanations

By the **causal variables**:



Many types of explanations

By the **completeness**: selective, or comprehensive?



Many types of explanations

[Lombrozo, 2012] also mentioned a taxonomy that is particularly fruitful in psychology:

Why is this tire round?

Mechanistic: The manufacturing process makes their tires round.

Functional/teleological: Round tires can reduce the friction.

Formal/categorical: It's round because it is a tire.

…

Tanya Lombrozo. 2012. Explanation and Abductive Inference. In Keith J. Holyoak and Robert G. Morrison, editors, *The Oxford Handbook of Thinking and Reasoning*, pages 260–276. Oxford University Press, 1st ed.

Many types of explanations

[Chen et al., 2023] grouped many XAI methods by “what is explained”:

> A horse walks [MASK] a river.
> My language model predicts the filled in word to be “across”. Why?

- Explain the **model’s decision boundary**
  - Why the model predicts “across” instead of “along”, etc.
- Explain the **task’s decision boundary**
  - Why “across” is the answer, instead of “along”.
- Explain the **errors of the predictors**.
  - If another model makes an incorrect prediction.

Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. 2023. Machine Explanations and Human Understanding. arXiv:2202.04092 [cs].

Many types of explanations

Are the explanations generated with the prediction, or after?

“This movie is excellent.” -> “Positive sentiment”

With-prediction explanation: e.g., feature attributions of *the* sentiment analysis model.

Post-hoc explanation: e.g., an external explainer tries to explain the prediction.

Many types of explanations

By the methods to find these explanations

- Prompt-based
- Influence functions
- Causal mediation
- Mechanistic interpretability methods
- Many others

This tutorial will approximately follow this taxonomy.

Outline of the tutorial

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

This section

Why study explanations?

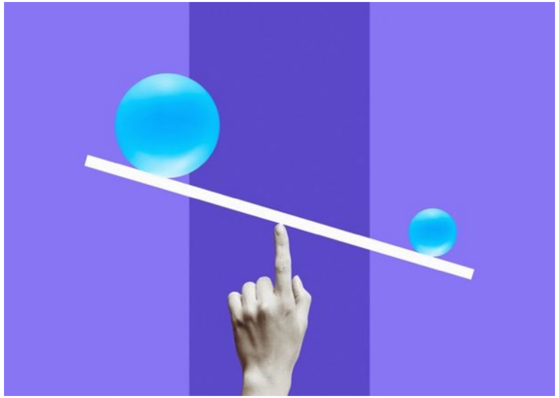
- Explanation is widely used in daily lives.
- Explanation can help people understand (& potentially improve) models.
- Large language models introduce unique challenges and opportunities for explanations.

Goals of this tutorial:

- **Review** these challenges and opportunities,
- **Connect** the lines of research previously studied by different research groups,
- **Spark** thoughts for new research directions.

Desirable properties of explanations

Following slides present an incomplete list.



Good explanations should be faithful

To explain a model's prediction:

"A **faithful** interpretation is one that accurately represents the reasoning process behind the model's prediction." [Jacovi & Goldberg, 2020]

"This is often considered the most fundamental requirement for any explanation, and sometimes used interchangeably with the term 'interpretability'." [Lyu et al, 2023]

Plausibility and faithfulness are two desirable (but different) properties in explanations. [Wiegreffe and Pinter, 2019]

Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? (Jacovi & Goldberg, ACL 2020)
Towards Faithful Model Explanation in NLP: A Survey. (Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2023)
Attention is not not Explanation (Wiegreffe & Pinter, EMNLP-IJCNLP 2019)

Good explanations should be plausible

An explanation is considered plausible if it is coherent with human reasoning and understanding. [Agrawal et al, 2024]

Plausibility is also referred to as **persuasiveness** or **understandability**.

An explanation might be plausible but not faithful. Currently, many explanations are more plausible than faithful.

Example of faithful, but not plausible explanation: a copy of model weights. [Jacovi & Goldberg, 2020]

Agarwal, C., Tanneru, S. H., & Lakkaraju, H. (2024). Faithfulness vs. Plausibility: On the (Un) Reliability of Explanations from Large Language Models. *arXiv preprint arXiv:2402.04614*.

Good explanations should be informative

Hi prof, I have just finished this paper. Which venue do you think would best suit it?

NAACL, because its deadline is just 3 days away, and it will be in Mexico, not far from here.

NAACL, because it is a top NLP conference.

Which explanation is more informative?

This example is modified from: Zachary C. Lipton. 2017. The Mythos of Model Interpretability. arXiv:1606.03490 [cs, stat].

Good explanations should be useful

Tom holds a copper block by hand and heats it on fire. Which of the following is more likely?

A. His fingers feel warm. B. His fingers feel burnt.

A (92.93%)

Copper is a good thermal conductor. Tom holds a copper block by hand and heats it on fire. Which of the following is more likely?

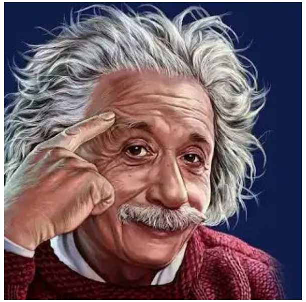
A. His fingers feel warm. B. His fingers feel burnt.

B (90.05%)

This example is modified from the e-CARE dataset. Probabilities are computed by GPT-3.5-Turbo model, as of April 19, 2024. This example is about utility to models – explanation should also be useful to humans.

Good explanations should be simple

“Everything should be made as simple as possible, but no simpler.”



“Everything should be made as simple as possible, but no simpler” might, says Calaprice, be a compressed version of lines from a 1933 lecture by Einstein: “It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.”

Other desirable properties

Completeness: All necessary information is covered.

Stability: The explanation remains consistent for similar cases I ask about.

Interactivity: The explanation can answer my follow-up questions.

Personalization: The explanation is tailored to my needs.

…

For a more detailed list please refer to [Liao et al., 2022]

Q. Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 10(1):147–159.

How is explanation different for LLMs?

Challenges

Lack of transparency

Computation: time, memory

Unreliable

Trust

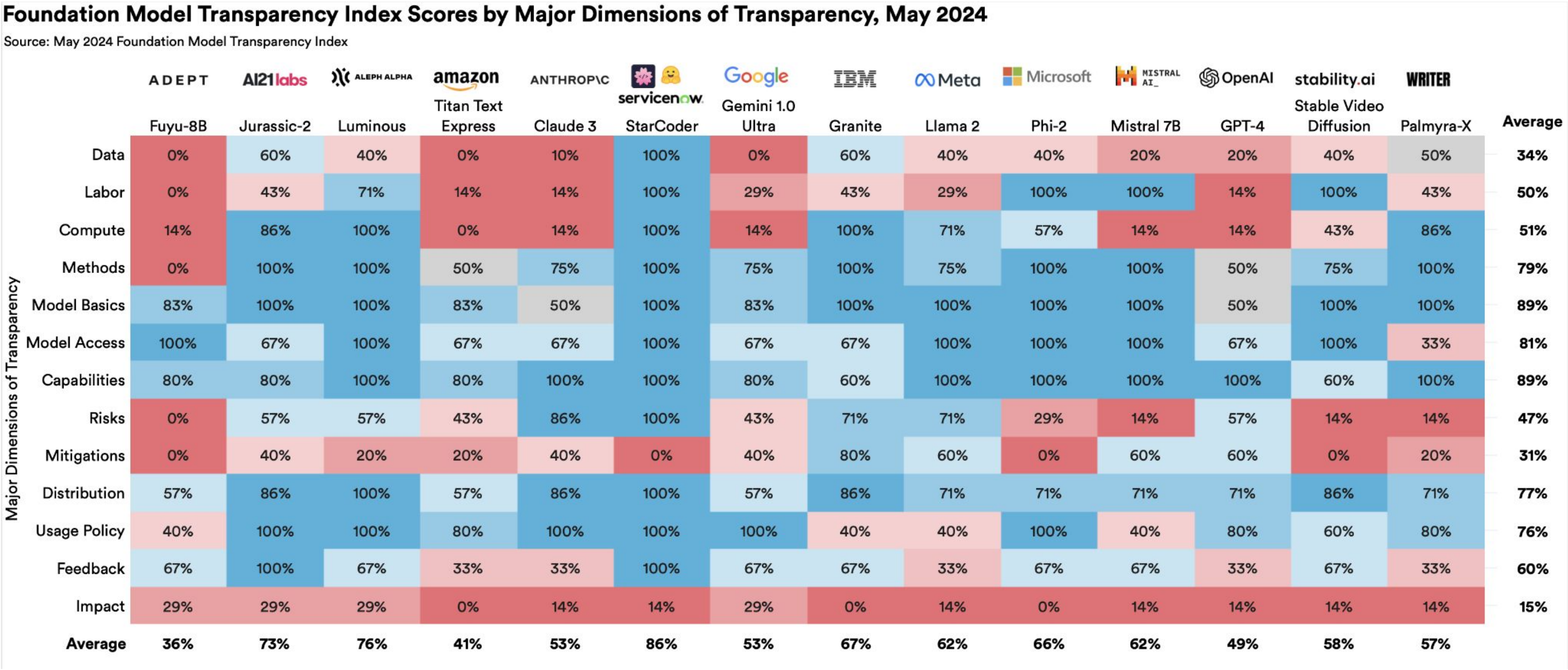
Evaluation

Opportunities

“This scaling also spurs innovation in interpretability techniques and offers richer insights into model behavior”

Emerging abilities

Challenge: Lack of transparency



Foundation Model Transparency Index Scores by Major Dimensions of Transparency, May 2024

Source: May 2024 Foundation Model Transparency Index

Major Dimensions of Transparency	ADEPT Fuyu-8B	AI21labs Jurassic-2	ALEPH ALPHA Luminous	amazon Titan Text Express	ANTHROP\C Claude 3	servicenow StarCoder	Google Gemini 1.0 Ultra	IBM Granite	Meta Llama 2	Microsoft Phi-2	MISTRAL AI_ Mistral 7B	OpenAI GPT-4	stability.ai Stable Video Diffusion	WRITER Palmyra-X	Average
Data	0%	60%	40%	0%	10%	100%	0%	60%	40%	40%	20%	20%	40%	50%	**34%**
Labor	0%	43%	71%	14%	14%	100%	29%	43%	29%	100%	100%	14%	100%	43%	**50%**
Compute	14%	86%	100%	0%	14%	100%	14%	100%	71%	57%	14%	14%	43%	86%	**51%**
Methods	0%	100%	100%	50%	75%	100%	75%	100%	75%	100%	100%	50%	75%	100%	**79%**
Model Basics	83%	100%	100%	83%	50%	100%	83%	100%	100%	100%	100%	50%	100%	100%	**89%**
Model Access	100%	67%	100%	67%	67%	100%	67%	67%	100%	100%	100%	67%	100%	33%	**81%**
Capabilities	80%	80%	100%	80%	100%	100%	80%	60%	100%	100%	100%	100%	60%	100%	**89%**
Risks	0%	57%	57%	43%	86%	100%	43%	71%	71%	29%	14%	57%	14%	14%	**47%**
Mitigations	0%	40%	20%	20%	40%	0%	40%	80%	60%	0%	60%	60%	0%	20%	**31%**
Distribution	57%	86%	100%	57%	86%	100%	57%	86%	71%	71%	71%	71%	86%	71%	**77%**
Usage Policy	40%	100%	100%	80%	100%	100%	100%	40%	40%	100%	40%	80%	60%	80%	**76%**
Feedback	67%	100%	67%	33%	33%	100%	67%	67%	33%	67%	67%	33%	67%	33%	**60%**
Impact	29%	29%	29%	0%	14%	14%	29%	0%	14%	0%	14%	14%	14%	14%	**15%**
Average	**36%**	**73%**	**76%**	**41%**	**53%**	**86%**	**53%**	**67%**	**62%**	**66%**	**62%**	**49%**	**58%**	**57%**	

Challenge: Computation

- Memory
  - Model can’t fit the memory.
  - Data can’t fit.
  - Computation results can’t fit.
  - GPU/GPUs with large memory are expensive.
- Time
  - Datasets are large.
  - Algorithms are complicated.
  - Trading off time for memory.

Challenge: Unstable

Explanations generated by the LLMs may be inconsistent, or unfactual. [Ye & Durrett, 2022]

Inconsistent: Explanations do not entail the models' predictions.

Unfactual: Explanations are not factually grounded in the input.

Language models don't always say what they think [Turpin et al., 2023]

LLMs cannot self-correct reasoning yet [Huang et al., 2024]

Challenge: Over-reliance

[Vasconcelos et al., 2022] “Explanations can reduce overreliance on AI systems during decision-making“

[Chen et al., 2023]: “Empirical studies have not found consistent evidence of explanations’ effectiveness and, on the contrary, suggest that they can increase overreliance when the AI system is wrong.” [Bansal et al., 2021][Poursabzi-Sangdeh et al., 2021][Wang and Yin, 2021][Zhang et al., 2020]

[Ajwani et al., 2024]: “LLM-generated black-box explanations can be adversarially helpful”

Challenge: Hard to evaluate

What to evaluate?

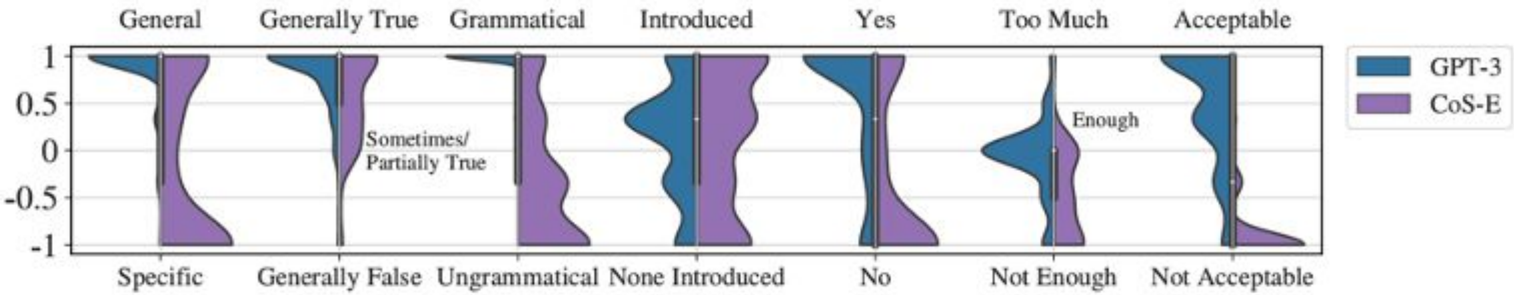
- There are many desirable properties.
  - Each property is hard to evaluate.
- Potential trade-off between the properties.
  - Plausibility vs faithful?
  - Informativeness vs simplicity?
  - …
  - Btw: they might not exactly be trade-offs.
  - Preference depends on the actual scenarios.

How to evaluate?

- The evaluations are specific to the explanation methods.
- More about evaluations in the following sections.

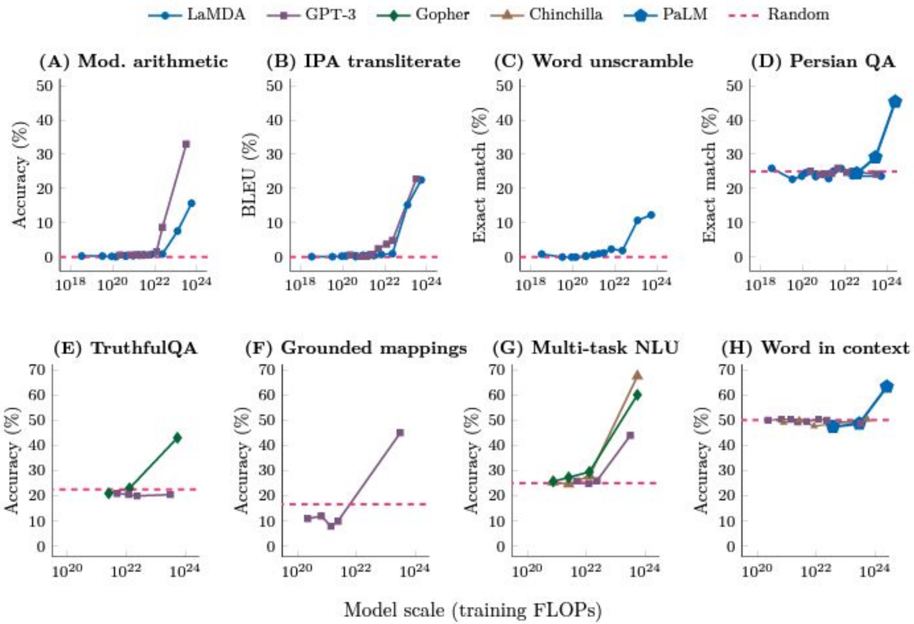
Opportunity: Strong capabilities

“In a head-to-head comparison, crowd-workers often prefer explanations generated by GPT-3 to crowdsourced explanations in existing datasets.” [Wiegreffe et al., 2022]



Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing Human-AI Collaboration for Generating Free-Text Explanations. In *NAACL*.

Opportunity: Emerging capabilities



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. arXiv:2206.07682 [cs].

Opportunity: Big data

Models are trained on unprecedented scales of datasets.

- Novel analysis targets.
- Novel analysis tools.

Outline of the tutorial

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

← This section

Summary of this section

- What is explanation, and why study explanation.
- Taxonomizations of explanations.
  - The remaining sections of this tutorial will follow a taxonomy by methods.
- Desired properties of explanations.
  - Faithfulness, plausibility, informative, useful, simple, etc.
- Challenges of explanations in the LLM era.
  - Compute and data
  - Unreliability
  - Trust
  - Difficulties in evaluation
  - Access
- Opportunities of explanations in the LLM era.
  - Strong capability
  - Massive data and knowledge